Project EC2: Process and Methods

Confiance.ai Taxonomy V2

contact@confiance-ai.fr | www.confiance.ai

Contents

Chapter A

Introduction

A.1. Objectives

This report proposes an operational definition of trustworthy artificial intelligence to be adopted in the context of the Grand challenge "Confiance.ai", covering to the overall life cycle of an AI-based critical system. This operational definition is constituted by a concise taxonomy and a list of keywords that characterize the core domains of the trustworthy artificial intelligence including the following fields : AI Engineering, Data Engineering, Knowledge Engineering, Algorithm Engineering, Software and System Engineering, Safety Engineering, Human factor and Cognitive Engineering. Thus, this report proposes a taxonomy for artificial intelligence (AI) and a list of related keywords, fixing concepts dealt with in the Confiance.ai program.

Based on the observations during the batch 2 and the first version of the Trustworthiness AI taxonomy, we will now suggest a new framework for mapping the Confiance.ai landscape that tries to cover the goals of:

- Covering all relevant dimensions of trustworthiness AI;
- The categories per dimensions should be mutually exclusive as well as collectively exhaustive;
- Technology-related dimensions should clearly focus on core AI technologies;
- Ability to map related engineering standards.

This taxonomy supports work to build a common digital and documented reference framework to harmonize the design and in-service support (including monitoring and maintenance) activities during the overall life-cycle of a critical AI-based system and to support the Confiance.ai methodology including requirements and recommendations.

A.2. Methodology

The various definitions have been collected through all the different states of the art realized within the Confiance.ai program. In most cases, the definitions were taken from external literature by the working group in charge of the state of the art; in some few cases, literature does not offer a definition adapted to the scope of the program, and the working group coined a new one.

This collection emanating from Confiance.ai's states of the art has been enriched in this present document with definitions coming from European and worldwide standardization (e.g. CEN, CENELEC, IEC, ISO), European projects (the Franco-Canadian DEEL project, JRC Flagship on AI), scientific publications or working groups such as the HLEG (High-Level Expert Group on Artificial Intelligence) or the AI Safety Landscape initiative (`https://www.ai-safety.org/`).

In line with the two other pillars of the Grand National Challenge (and more specifically Pillar 3 dedicated to the French AI standardization strategy), this taxonomy will be updated during the overall duration of Confiance.ai, mainly to take into account the outcomes of all 7 projects (EC*) and also the outcomes of all batches. In addition, this taxonomy will be consolidated to define the trustworthy AI engineering ontology.

Moreover, some definitions could be duplicated because they are related to several engineering fields (algorithm engineering, software engineering, system engineering, cognitive engineering, cyber security, safety) or various AI paradigms (data engineering, knowledge engineering data-driven AI, knowledge-based AI).

A.3. Acronym

Acronym	Definition
AI	Artificial Intelligence
DL	Deep Learning
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
KBS	Knowledge-based System
KE	Knwoledge Engineering
KRR	Knowledge Representation and Reasoning
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SE	System Engineering

Contributors

	Name Organisation	Role
Agnès DELABORDE	LNE	Co-author
Juliette MATTIOLI	Thales	Resp & Co-author

Document Control

Revision	Date	Commentary	Author
V1.0	15/12/2021	1st version from Batch 1	A. Delaborde, J. Mattioli
V1.1	20/10/2022	1st release of the version V2	J. Mattioli
V2.0	27/12/2022	Document delivery of V2	J. Mattioli

A.4. Document references

This report covers the following expected contractual deliverables:

Ref. N°	Deliverable title
L2.1.1.2	The Confiance.ai Taxonomy (Version 2)
L2.1.2.1	Terminology on trustworthy AI jobs

Chapter B

Data, Information, Knowledge

B.1. Introduction

"It is impossible to make good decisions without relevant information".

According to [Ackoff, 1989], the content of the human mind can be classified into four categories: Data (D), Information (I), Knowledge (K), Wisdom (W), commonly called the DIKW pyramid. Data, Information and knowledge become decisional resources that must be managed smartly. But, data, facts and information are often used interchangeably with knowledge. Nevertheless, data represents the properties of objects and events.

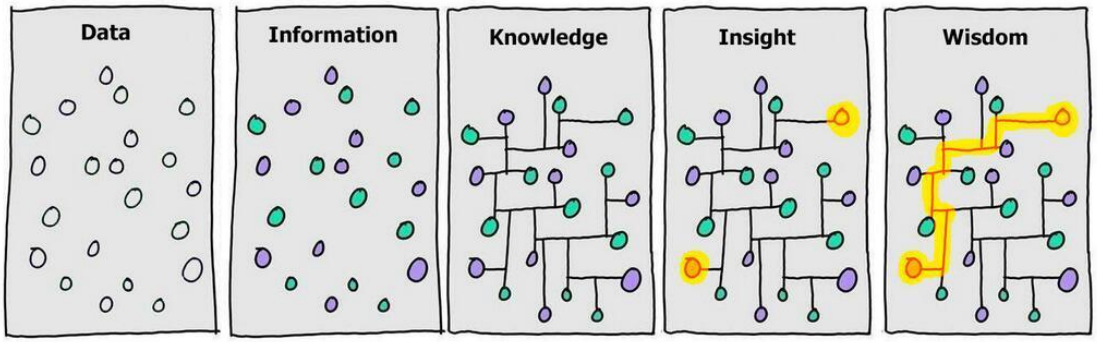

Figure B.1: Data, Information, Knowledge, Insight, Wisdom

In figure B.1, **Data** is represented by a series of random dots that could mean something – or nothing. Next comes **Information** where meaning or relationship are extracted from the raw data (indicated by colors associated to the dots). **Knowledge** is obtained when we are able to memorize the information, for example: standard multiplication tables or sunrise and sunset times in a given month. As we gain knowledge we begin to make sense of the data by drawing

connections between different pieces of information. However, it is at the **Insight** level where data becomes effectively useful. Insight is the ability to synthesize knowledge in order obtain a deep understanding of a problem. With insight comes the prospect of **Wisdom** – the ability to use insight to facilitate informed decision making.

This report focuses on data, information and knowledge.

Data are provided to the intelligent reasoning engines in order to give them a representation of the world. Once the data samples are acquired by the systems, they become information. This information is then interpreted by the intelligent systems, according to the context related to the acquisition of the data items, and the domain knowledge, among others. The domain knowledge is used to interpret a set of data items within a specific context.

Like data, **information** also represents the properties of objects and events, but in a more compact and useful form. The difference between data and information is functional, not structural. Information can be represented as descriptions, answers to questions that begin with such words as who, what, when, where, and how many.

Knowledge is conveyed by instructions, answers to how-to questions. Understanding is represented by explanations, answers to why questions. No understanding is possible without knowing the context in which the process of perception of a situation occurs.

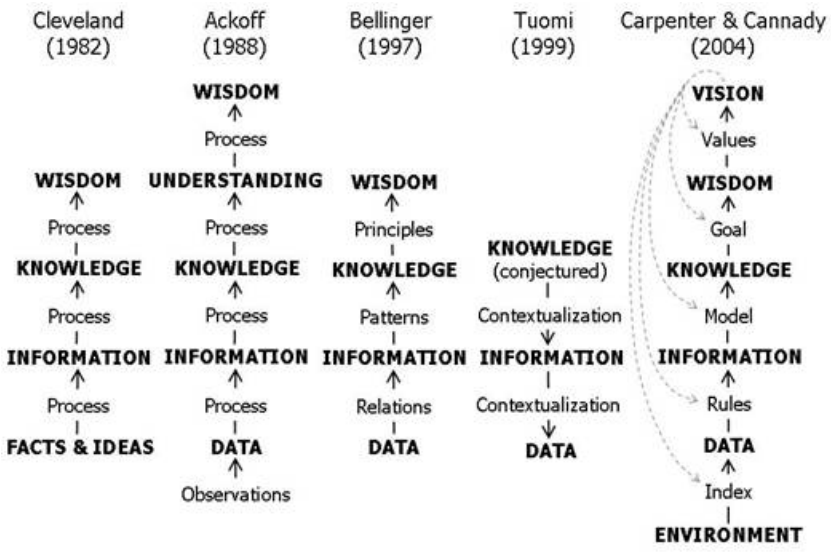

Figure B.2: Evolution of the DIKW model

Thus, data are the basic individual items of numeric information, garnered through observation; but in themselves, without context, they are devoid of information. Information is what is represented, and possibly amenable to analysis and interpretation, through data and the context

in which the data are assembled. Knowledge is the general understanding and awareness garnered from accumulated information, tempered by experience, enabling generalization to new contexts.

Information once analyzed, understood, and explained is knowledge, or foreknowledge (predictions or forecasts). Understanding information provides 1) a degree of comprehension of both the static and dynamic relationships of the objects of data, 2) the ability to model structures, and 3) past (and future) behavior of those objects. Knowledge includes both static content and dynamic processes. Thus, in creating knowledge the act of understanding information provides: 1) a degree of comprehension of both the static and dynamic relationships of the objects represented by the data, 2) the ability to model structures, and 3) past (and future) behavior of those objects. Knowledge includes both static content and dynamic processes.

B.2. First definitions

Big Data

- [Artificial Intelligence Roadmap, 2020] A discipline or set of methods that specializes in dealing with the analysis of very large amounts of data (more than terabytes), with a high velocity (high speed of data processing), from various sources (sensors, images, texts, etc.), and which might be unstructured (not standardized format).

Cognition

- [Stevenson, 2015] The mental action or process of acquiring knowledge and understanding through thought, experience, and the senses.

Cognitive science (cognivitism)

- [ISO/IEC 2382, 2015] Interdisciplinary knowledge field, whose stated objective is to discover the representational and computational capacities of the mind and their structural and functional representation in the brain.
- Cognitive science deals with the symbol-processing nature of cognition and encompasses disciplines as diverse as psychology, computer science, linguistics, anthropology, philosophy, education, mathematics, engineering, physiology, and neuroscience.

Connectionism (connectionist paradigm)

- [ISO/IEC DIS 22989, 2021a] Form of cognitive modelling that uses a network of interconnected units which generally are simple computational units.

Data

- [ISO/IEC 2382, 2015] Re-interpretable representation of information in a formalized manner suitable for communication, interpretation or processing. Data can be processed by

human or automatic means.
- [Ackoff, 1989] Data are provided to the intelligent reasoning engines in order to give them a representation of the world. Once the data samples are acquired by the systems, they become information. This information is then interpreted by the intelligent systems, according to the context related to the acquisition of the data items, and the domain knowledge, among others. The domain knowledge is used to interpret a set of data items within a specific context.

Data annotation

- [ISO/IEC DIS 22989, 2021a] Process of attaching a set of descriptive information to data without any change to that data.

Data augmentation

- [ISO/IEC DIS 22989, 2021a] Process of creating new data samples by manipulating the original data.

Data engineering

- The discipline that aims to organize, structure, trace and select data in such a way that its quality, availability, relevance and traceability can be guaranteed throughout the life cycle of the data.
- [ISO/IEC 20546, 2019] Discipline which is related to the engineering aspects of systems, processing, models and management of data, including but not limited to big data.

Data mining

- [ISO/IEC DIS 22989, 2021a] Computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing it, and summarizing potential relationships and impacts.

Data quality

- [ISO/IEC 25024, 2015] Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions.
- [Mamalet et al., 2021] The extent to which data are free of defects and possess desired features

Data sampling

- [ISO/IEC DIS 22989, 2021a] Process to select a subset of data samples intended to present patterns and trends similar to that of the larger dataset being analyzed.

Data sciences

- [ISO/IEC 20546, 2019] Data science refers to the process for extracting knowledge from data – the approach can be either through exploration or by hypothesis testing. Data science refers to the complete data analytics lifecycle where data analytics is understood.
- [EASA, 2020] A broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the extraction of information and analysis to acquire knowledge from that information.

Data-driven AI

- [EASA, 2020] The data-driven approach focuses on building a system that can learn what is the appropriate answer based on having trained on a large number of examples.

Dataset

- [Branco et al., 2008] An aggregation of data, typically spawning more than one physical file,that are processed together and serve collectively as input or output of a computation or data acquisition process.

Descriptive analytics

- [Logility, 2021] Descriptive Analytics use data aggregation and data mining to provide insight into the past and answer: “What has happened?”

Domain knowledge

- [ISO/IEC 2382, 2015] Knowledge accumulated in a particular domain.

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Information

- [ISO 9000, 2015] Meaningful data. [data being "facts about an object"]

Information

- [Ackoff, 1989] Information represents the properties of objects and events, but in a more compact and useful form. The difference between data and information is functional, not structural. Information can be represented as descriptions, answers to questions that begin with such words as who, what, when, where, and how many.

Information (information processing)

- [ISO/IEC 2382, 2015] Knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning.

Information (information theory)

- [ISO/IEC 2382, 2015] Knowledge which reduces or removes uncertainty about the occurrence of a specific event from a given set of possible events. In information theory, the concept "event" is to be understood as used in the theory of probability. For instance, an event may be: the presence of a specific element in a given set of elements; the occurrence of a specific character or word in a given message or in a given position of a message; any one of the distinct results an experiment may yield.

Information analysis

- [ISO/IEC 2382, 2015] Systematic investigation of information and its flow in a real or planned system.

Intelligence

- [Shhab et al., 2005] The capability of learning, understanding and finding solutions for problems in a specific domain.

Knowledge

- The theoretical and practical comprehension of a certain domain, that supports making decisions.
- [Ackoff, 1989] Knowledge is conveyed by instructions, answers to how-to questions. Understanding is represented by explanations, answers to why questions. No understanding is possible without knowing the context in which the process of perception of a situation occurs.

Knowledge acquisition

- [ISO/IEC 2382, 2015] Process of locating, collecting, and refining knowledge and converting it into a form that can be further processed by a knowledge-based system. Knowledge acquisition normally implies the intervention of a knowledge engineer, but it is also an important component of machine learning.

Knowledge base

- [ISO/IEC 2382, 2015] Database that contains inference rules and information about human experience and expertise in a domain. In self-improving systems, the knowledge base additionally contains information resulting from the solution of previously encountered problems.

Knowledge engineering

- [ISO/IEC 2382, 2015] Discipline concerned with acquiring knowledge from domain experts and other knowledge sources and incorporating it into a knowledge base. The term "knowledge engineering" sometimes refers particularly to the art of designing, building, and maintaining expert systems and other knowledge-based systems.

Knowledge graph

- [ICF, 2018] Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities.

Knowledge representation

- [ISO/IEC 2382, 2015] Process or result of encoding and storing knowledge in a knowledge base

Knowledge-based methods

- [Scharei et al., 2020] Knowledge-based methods are based on ontologies and huge databases of notions, information, and rules.

Knowledge-based system

- [ISO/IEC 2382, 2015] Information processing system that provides for solving problems in a particular domain or application area by drawing inferences from a knowledge base. The term "knowledge-based system" is sometimes used synonymously with "expert system", which is usually restricted to expert knowledge. Some knowledge-based systems have learning capabilities.

Label (organisation of data)

- [ISO/IEC 2382, 2015] Identifier that is attached to a set of data elements.

Life cycle

- [ISO/IEC/IEEE 15288, 2015] The evolution of a system, product, service, project or other human-made entity from conception through retirement. A life cycle can be described using an abstract functional model that represents the conceptualization of a need for the system, its realization, utilization, evolution and disposal.
- [ISO/IEC DIS 22989, 2021a] Evolution of a system, product, service, project or other human-made entity, from conception through retirement.

Model

- [Object Management Group, 2010] A selective representation of some system whose form and content are chosen based on a specific set of concerns. The model is related to the system by an explicit or implicit mapping.
- [of Defense, 1996] A physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process.
- [Friedenthal et al., 2007] A representation of one or more concepts that may be realized in the physical world.
- [Daneshjo, 2012] A simplified representation of a system at some particular point in time or space intended to promote understanding of the real system.
- [EUROCAE ED 218, 2012] Abstract representation of a given set of aspects of a system/subsystem.

Operational concept

- [ISO/IEC/IEEE 15288, 2015] Verbal and graphic statement of an organization's assumptions or intent in regard to an operation or series of operations of a system or a related set of systems.

Operational Design Domain (ODD)

- [SAE J3016, 2018] Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristic.

Pattern (artificial intelligence)

- [ISO/IEC 2382, 2015] Set of features and their relationships used to recognize an entity within a given context. These features may include a geometrical shape, a sound, a picture, a signal, or text.

Predictive analytics

- [Logility, 2021] Prescriptive Analytics use optimization and simulation algorithms such as knowledge based AI to advise on possible outcomes and answer: "What should we do?"

Prescriptive analytics

- [Logility, 2021] Predictive Analytics use statistical models and forecasting techniques such as ML to understand the future and answer: "What could happen?"

Probabilistic methods

- [Scharei et al., 2020] Probabilistic methods allow acting in incomplete information scenarios.

Reasoning (domain)

- [Samoili et al., 2020] The domain of reasoning tackles the way machines transform data into knowledge, or infer facts from data. Several classifications address knowledge representation and automated reasoning as a field of AI, to describe the process of justifying (reasoning) the available data and information, provide solutions and represent them efficiently, based on a set of symbolic rules

Uncertainty

- [ISO/IEC 2382, 2015] Condition appearing when a value cannot be determined during consultation, or a fact or a rule in the knowledge base remains in doubt.

Chapter C

Artificial Intelligence

C.1. Introduction

AI techniques and sub-disciplines can be grouped under two big paradigms [HLEG, 2018] regarding the systems' capabilities (see Fig.C.1): (i) reasoning and decision making, (ii) and learning and perception. The first group of capabilities includes the transformation of data into knowledge, by transforming real world information into something understandable and usable by machines, and making decisions following an organized path of planning, solution searching and optimization. This knowledge-based AI domain covers Knowledge representation & reasoning (usually making use of symbolic rules to represent and infer knowledge) and Planning (including Planning & Scheduling, Searching, and Optimization). The second paradigm develops in absence of symbolic rules, and involves learning -meaning the extraction of information, and problem solving based on structured or unstructured perceived data (written and oral language, image, sound, etc.), adaptation and reaction to changes, behavioral prediction, etc. This data-driven AI domain covers AI sub-fields related to learning, communication and perception, such as Machine learning, Natural language processing, and Computer vision.

In the following, we distinguish between three major paradigms in artificial intelligence: Data-driven AI, Knowledge-based AI and Hybrid AI (see Fig.C.2)

Knowledge-based AI is the branch of AI that simulates the mind's symbolic processing by attempting to explicitly represent human knowledge in a declarative form (i.e. using facts and rules). Thus, to be successful in producing human-like intelligence, it is necessary to translate implicit or procedural knowledge into an explicit form using symbols and rules for their manipulation. Therefore, symbolic approaches are based upon a syntax that is endowed with formal semantics (meaning) useful for properties expression and verification. They have been successfully applied to increase the system's trustworthiness in different application domains like health care, automotive, aeronautics, railway [Hofer-Schmitz and Stojanović, 2020].

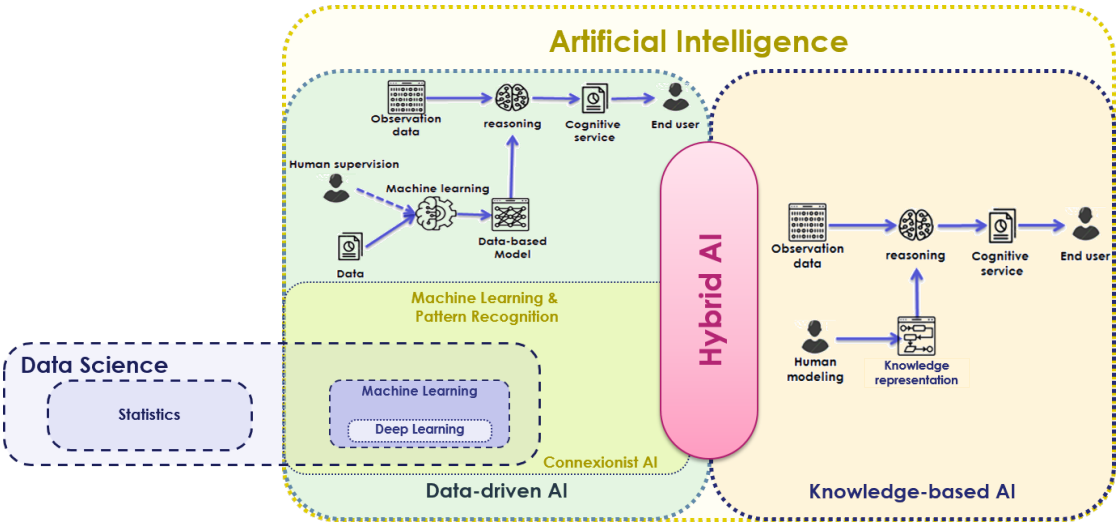

Figure C.1: Data-driven AI, Knowledge-based AI and Hybrid AI paradigms illustrated with some techniques

Contrary to Knowledge-based AI, **data-driven AI** is not built upon an explicit representation of human expertise: the behavior is learnt from examples/data. Usually, we distinguish between two categories of problems that can be addressed:

- **Classification problem**: find a function which maps an input to a discrete output, also called label or class. This problem is referred to as binary classification when there are two classes, and multi-class classification when there are more than two classes. Spam detection is an example of binary classification problem: find a function which determines whether an email is a spam or not. Image recognition is a typical example of multi-class detection, when a user is interested in getting what is represented in an image (e.g., street sign recognition).
- **Regression problem**: find a function which maps an input to a continuous output. A typical example is to "reproduce" an unknown function when given a large collection of inputs and corresponding outputs (e.g., price prediction of a square meter for a given area).

In the data-driven AI paradigm, **connectionist AI** algorithms are based upon a statistical or probabilistic model which is tightly coupled to data sets which are first used for model training, and once tuned, for performance evaluation. The model is often structured as a set of nodes defined by multi-value functions or random variables. The nodes are interconnected, and the links can be randomly weighted by values influencing nodes inputs/outputs [Kasabov, 2012].

More recently [Sun, 2015], **hybrid** approaches integrating knowledge-based and connectionist paradigms have been proposed. Hybrid approaches aim to profit of salient features of both symbolic and connectionist leaving out any potential concurrence [Foggia et al., 2001].

In certain cases, the complementary between techniques even leads to overcome certain limita-

tions of each other.

On one side, data-driven AI approaches successfully characterize and capture the salient traits of the data sets. However, being the connectionist models heuristic and agnostic of typical notion-encapsulation archetypes, they lack argumentation necessary for explainability. On the other side, symbolic approaches introduce a semantic layer aligned with the notion-encapsulation archetype, which is amenable for expressing domain knowledge and concerns useful for validation and argumentation.

C.1.1 Data-driven AI

Current interest in artificial intelligence is almost entirely focused on data-driven AI.

The reasons are easy to understand. Cheap data storage, fast processors and advancements in neural net algorithms and other data-centric techniques have made it possible to extract huge value out of data. We can build systems that can predict what will happen next based on what they have seen so far, very efficiently. Their performance is at times even better than that of a human being.

Connectionist AI gets its name from the typical network topology that most of the algorithms in this class employ. The most popular technique in this category is the Neural Network (NN). This consists of multiple layers of nodes, called neurons, that process some input signals, combine them together with some weight coefficients, and squash them to be fed to the next layer. Support Vector Machines (SVMs)[1] also fall under the connectionist category. Often used in the context of pattern recognition, classification, clustering or perception, connectionist AI such as **machine learning** seeks to capture tacit knowledge - knowledge which is difficult or impractical to explicitly define - through statistical approaches by inferring the inherent structure from a set of examples (data) which can be used for mapping new examples. ANNs (Artificial Neural Networks) come in various shapes and sizes, including Convolution Neural Networks (successful for image recognition and bitmap classification), and Long Short-term Memory Networks (typically applied for time series analysis or problems where time is an important feature). Deep learning is also essentially synonymous with Artificial Neural Networks.

Supervised learning methods are done using ground truth; unsupervised learning methods do not use explicitly provided labels. Using machine learning, buyers and suppliers could collaborate more effectively and reduce stock-outs, improve forecast accuracy, and meet or beat more customer delivery dates.

[1] In machine learning, support-vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

C.1.2 Knowledge-based AI

Knowledge-based AI can be defined as a problem-solving system such as resource allocation, planning, or (multi-criteria) decision making, utilizing symbolic model and knowledge through a reasoning process. A knowledge-based AI system works by carrying out a series of logic-like reasoning steps over language-like representations. The representations are typically propositional in character, and assert that certain relations hold between certain objects, while each reasoning step computes a further set of relations that follow from those already established, according to a formally specified set of inference rules. A particular feature is based on a separation between the knowledge, which can be represented by various approaches such as rules or constraints and the inference or reasoning algorithm which uses the knowledge base to build a conclusion. Knowledge-based AI uses a set of defined knowledge (or business rules) to derive and manipulate data. The integration of such an approach will, for example, bring about automatic scheduling of jobs through a production facility based on rules and parameters set by the production manager.

C.1.3 Hybrid AI

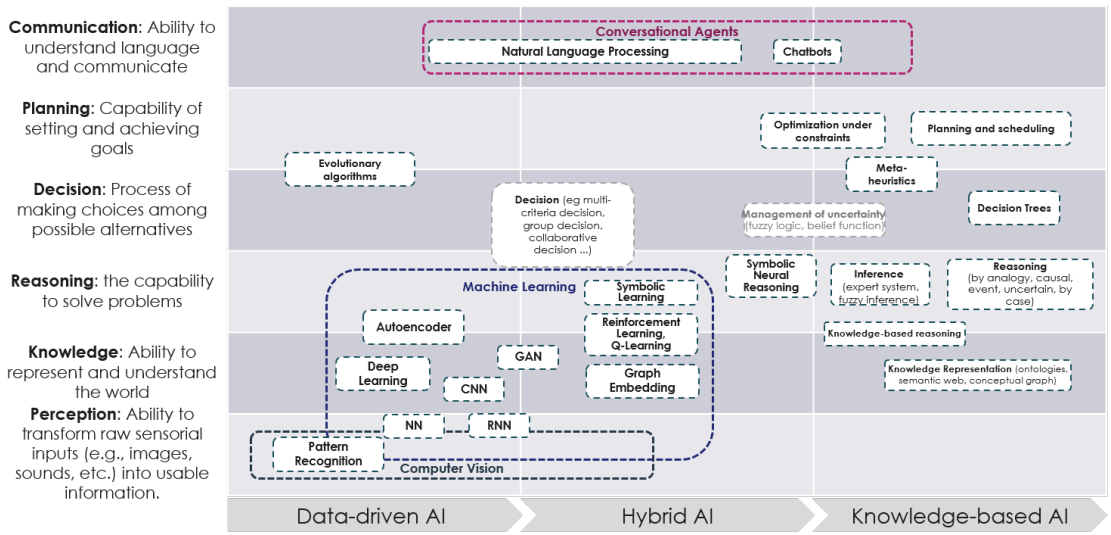

Figure C.2: AI Algorithm paradigm illustrated with some AI technics

Over the past ten years, data-driven AI has become established as one of the most impactful research areas within AI. Notwithstanding its undeniable success, critics have recently drawn attention to a number of shortcomings in contemporary deep learning. Neural networks (NN) require large volumes of training data to be effective. They are prone to fail disastrously when exposed to data outside the distribution they were trained on. Data-driven AI has an inherent "black box" nature. The computations carried out by successive layers rarely correspond to humanly comprehensible reasoning steps, and the intermediate vectors of activations they generate usually lack a humanly comprehensible semantics. In simple terms, you may not always know why the system made the choices / recommendations / predictions it made, which is a roadblock for

deploying AI within a critical system. This necessitated an all-inclusive approach called hybrid AI – taking the best of what knowledge-based AI and physics-based modeling offered and combining it with the capabilities of machine learning. On the other hand, an important limitation of knowledge-based AI relates to the so called "symbol grounding problem" [Harnad, 1990], and concerns the extent to which its representational elements are handcrafted rather than learned from data (e.g. from sensory input). By contrast, one of the strengths of deep learning is its ability to discover features in high-dimensional data with little or no human intervention. Significantly, the shortcomings of data-driven AI align with the strengths of Knowledge-based AI, which suggests the time is right for a hybridization (see table C.1 and figure C.2).

	Strengths	Weaknesses
Data-driven AI	As long as it does not aspire to be a symbol system, Data-driven AI has the advantage of not being subject to the symbol grounding problem. It applies the same small family of algorithms to many problems, whereas symbolic-AI is more a methodology rather than an algorithm relying on problem-specific symbolic rules. Its architecture seems more brain-like than a Turing machine or a digital computer. It is suited to the learning of patterns from data.	Because they are not symbol systems, do not have the systematic semantic properties that many cognitive phenomena appear to have. Not every problem amounts to learning. Some cognitive activities may call for problem specific rules, symbol manipulation, and standard computation. It may (like toy models) camouflage deeper performance limitations.
Knowledge-based AI	Symbols have the computing power of Turing machines and the systematic properties of a formal syntax that is semantically interpretable. All computable functions are equivalent to a computational state in a Turing machine	It is subject to the symbol grounding problem. Turing power is too general. The solutions to AI's many toy problems do not give rise to common principles of cognition but to a vast variety of ad-hoc symbolic strategies.

Table C.1: Data-driven AI versus Knowledge-based AI

C.2. General definitions

Abductive inference

- [ISO/IEC 2382-28, 1995] Inference from particular facts to plausible explanations of these facts.

Agent

- [ISO/IEC DIS 22989, 2021a] Automated entity that perceives its environment and takes actions to achieve its goals.

AI explainability

- [EASA, 2021] The AI explainability building block deals with the capability to provide the human with understandable and relevant information on how an AI/ML application is coming to its results.

AI safety risk

- [EASA, 2021] The AI safety risk mitigation building block considers that we may not always be able to open the AI black box to the extent required and that the safety risk may need to be addressed to deal with the inherent uncertainty of AI.

Artificial Intelligence (AI)

- [Brundage et al., 2020] Any digital system capable of performing tasks commonly thought to require intelligence, with these tasks often being learned from data and/or experience.
- [ISO/IEC/IEEE 24765, 2017] The branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self improvement.
- [ISO/IEC TR 24028, 2020] Capability of an engineered system to acquire, process and apply knowledge and skills. Knowledge are facts, information and skills acquired through experience or education.
- [JRC, 2018] AI is a generic term that refers to any machine or algorithm that is capable of observing its environment, learning, and based on the knowledge and experience gained, taking intelligent action or proposing decisions.
- [Statec, 2021] Artificial intelligence refers to systems that use technologies such as: text mining, computer vision, speech recognition, natural language generation, machine learning, deep learning to gather and/or use data to predict, recommend or decide, with varying levels of autonomy, the best action to achieve specific goals.
- [European Commission, 2018] Artificial Intelligence refers to systems that display intelligent behavior by analysing their environment and taking action Ñ with some degree of autonomy Ñ to achieve specific goals

- [Russell and Norvig, 2016] AI is the study of methods allowing to the computer to behave intelligently AI includes tasks as learning, reasoning, planning, perception, language comprehension and robotics these technologies aim to achieve with computer science cognitive tasks that are traditionally achieved by human beings.
- [JORF, 2018] A theoretical and practical interdisciplinary field, with the objective of understanding the cognitive and thinking mechanisms, and their imitation by a material and software device, for assistance or substitution purposes of human activities.
- [The Danish Government, 2019] Artificial intelligence is systems based on algorithms (mathematical formulae) that, by analysing and identifying patterns in data, can identify the most appropriate solution. The vast majority of these systems perform specific tasks in limited areas, *e. g.* control, prediction and guidance. The technology can be designed to adapt its behavior by observing how the environment is influenced by previous actions.

Artificial Intelligence development

- [Brundage et al., 2020] AI development refers to the process of researching, designing, testing, deploying, or monitoring AI.

Artificial Intelligence System

- [High-Level Expert Group on Artificial Intelligence, 2019] AI systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analysing how the environment is affected by their previous actions
- [ISO/IEC TR 24028, 202O] Working definition of AI system: any system (whether a product or a service) that uses AI. There are many different kinds of AI systems. Some are implemented completely in software, while others are mostly implemented in hardware (e.g. robots).
- [OECD - AIGO, 2019] An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. It uses machine and/or human-based inputs to perceive real and/or virtual environments; abstract such perceptions into models (in an automated manner e. g. with ML or manually); and use model inference to formulate options for information or action. AI systems are designed to operate with varying levels of autonomy.
- Computerized system that uses cognition to understand information and solve problems.
  - Note 1: [ISO/TR 23482-1, 2020] "Information technology –Vocabulary" defines AI as "an interdisciplinary field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning".
  - Note 2: In computer science AI research is defined as the study of "intelligent

agents": any device that perceives its environment and takes actions to achieve it goals.
  - Note 3: This includes pattern recognition and the application of machine learning and related techniques.
  - Note 4: Artificial Intelligence is the whole idea and concepts of machines being able to carry out tasks in a way that mimics the human intelligence and would be considered "smart".
- [ISO/IEC TR 24028, 2020] System using AI.

Audio processing

- [Samoili et al., 2020] Audio processing refers to AI systems allowing the perception or generation (synthesis) of audio signals, including speech, but also other sound material (e.g. environmental sounds, music). Speech or voice recognition, audio processing or sound technologies are also often proposed to be archived as an AI subdivision.

Connectionism (connectionist paradigm)

- [ISO/IEC DIS 22989, 2021a] Form of cognitive modelling that uses a network of interconnected units which generally are simple computational units.

Data mining

- [ISO/IEC DIS 22989, 2021a] Computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing it, and summarizing potential relationships and impacts.

Data sciences

- [ISO/IEC 20546, 2019] Data science refers to the process for extracting knowledge from data Ð the approach can be either through exploration or by hypothesis testing. Data science refers to the complete data analytics lifecycle where data analytics is understood.
- [EASA, 2020] A broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the extraction of information and analysis to acquire knowledge from that information.

Data-driven AI

- [EASA, 2020] The data-driven approach focuses on building a system that can learn what is the appropriate answer based on having trained on a large number of examples.

Deep learning

- [ISO/IEC TR 29119-11, 2020] Approach to creating rich hierarchical representations through the training of neural networks with one or more hidden layers.

- [EASA, 2020] A specific type of machine learning based on the use of large (deep) neural networks to learn abstract representations of the input data through the use of multiple layers.

Descriptive Analytics

- [Logility, 2021] Descriptive Analytics use data aggregation and data mining to provide insight into the past and answer: What has happened?

Domain (artificial intelligence)

- [ISO/IEC 2382, 2015] Specific field of knowledge or expertise.

Domain knowledge

- [ISO/IEC 2382, 2015] Knowledge accumulated in a particular domain.

Domain model

- [ISO/IEC 2382, 2015] Model of a specific field of knowledge or expertise.

Expert system

- [ISO/IEC 2382, 2015] Computer system that provides for expertly solving problems in a given field or application area by drawing inferences from a knowledge base developed from human expertise. The term "expert system" is sometimes used synonymously with "knowledge-based system", though it is usually taken to emphasize expert knowledge. Some expert systems are able to improve their knowledge base and develop new inference rules based on their experience with previous problems.

Explainability

- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.
- [Phillips et al., 2020] Explanation: Systems deliver accompanying evidence or reason(s) for all outputs.
  - Meaningful: Systems provide explanations that are understandable to individual users.
  - Explanation Accuracy: The explanation correctly reflects the systems process for generating the output.
  - Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.

- [Arrieta et al., 2020] Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.

Explainable Artificial Intelligence (XAI)

- [Arrieta et al., 2020] An explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Explainable model

- [Broniatowski et al., 2021] An explanation of a model result is a description of how a models outcome came to be.

Explanation facility

- [ISO/IEC 2382, 2015] Component of a knowledge-based system that explains how solutions were derived and justifies the steps used in reaching them.

Generative adversarial networks (GANs)

- [Techopedia, 2018b] A type of construct in neural network technology that is composed of two neural networks: a generative network, that generates samples, and a discriminative network, that tries to detect whether a sample is real or the result of the generator.

Genetic algorithm

- [ISO/IEC DIS 22989, 2021a] Algorithm which simulates natural selection by creating and evolving a population of individuals (solutions) for optimization problems.

Global Robustness

- [Mamalet et al., 2021] Ability of the system to perform the intended function in the presence of abnormal or unknown inputs

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Heuristic method

- [ISO/IEC 2382, 2015] Any exploratory method of solving problems in which an evaluation is made of the progress towards an acceptable final result using a series of approximate results, for example by a process of guided trial and error.

Heuristic rule

- [ISO/IEC 2382, 2015] Ad hoc rule written to formalize the knowledge and experience an expert uses to solve a problem.

Heuristic Search

- [ISO/IEC 2382, 2015] Search, based on experience and judgment, used to obtain acceptable results without guarantee of success.

Hierarchical Planning

- [ISO/IEC 2382, 2015] Planning that refines the vague parts of a plan into more detailed subplans by generating a hierarchical representation of it.

High-risk AI System

- [European Commission, 2021] AI systems that create a high risk to the health and safety or fundamental rights of natural persons.

Human Cognitive Bias

- [ISO/IEC TR 24027, 2021] Bias that occurs when humans are processing and interpreting information Note 1: human cognitive bias influences judgment and decision-making.

Inference

- [ISO/IEC 2382, 2015] Reasoning by which conclusions are derived from known premises.

Inference engine

- [ISO/IEC 2382, 2015] Component of an expert system that applies principles of reasoning to draw conclusions from representations of information stored in a knowledge base.

Intelligence

- [Shhab et al., 2005] The capability of learning, understanding and finding solutions for problems in a specific domain.

Interpretability

- [Arrieta et al., 2020] The ability to explain or to provide the meaning in understandable terms to a human.
- [Mamalet et al., 2021] Relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a human being. It is a basic requirement for an explanation.

Knowledge acquisition

- [ISO/IEC 2382, 2015] Process of locating, collecting, and refining knowledge and converting it into a form that can be further processed by a knowledge-based system. Knowledge acquisition normally implies the intervention of a knowledge engineer, but it is also an important component of machine learning.

Knowledge base

- [ISO/IEC 2382, 2015] Database that contains inference rules and information about human experience and expertise in a domain.
  In self-improving systems, the knowledge base additionally contains information resulting from the solution of previously encountered problems.

Knowledge Engineering

- [ISO/IEC 2382, 2015] Discipline concerned with acquiring knowledge from domain experts and other knowledge sources and incorporating it into a knowledge base.
  The term "knowledge engineering" sometimes refers particularly to the art of designing, building, and maintaining expert systems and other knowledge-based systems.

Knowledge graph

- [ICF, 2018] Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities.

Knowledge representation

- [ISO/IEC 2382, 2015] Process or result of encoding and storing knowledge in a knowledge base

Knowledge-based methods

- [Scharei et al., 2020] Knowledge-based methods are based on ontologies and huge databases of notions, information, and rules.

Knowledge-based system

- [ISO/IEC 2382, 2015] Information processing system that provides for solving problems in a particular domain or application area by drawing inferences from a knowledge base. The term "knowledge-based system" is sometimes used synonymously with "expert system", which is usually restricted to expert knowledge. Some knowledge-based systems have learning capabilities.

Learning

- [Scharei et al., 2020] Learning indicates incrementation of knowledge through experience.

Learning (machine learning)

- [ISO/IEC 2382, 2015] Process by which a biological or an automatic system gains knowledge or skills that it may use to improve its performance.

Learning (neural networks)

- [ISO/IEC 2382, 2015] Process by which a neural network improves its performance by adjustment of its parameters in response to a succession of input patterns. In general, learning consists in connection weights adjustment.

Learning by analogy (associative learning)

- [ISO/IEC 2382, 2015] Learning strategy that combines inductive learning and deductive learning so that inductions determine the common characteristics of concepts being compared or associated, and deductions derive from these characteristics the features expected of the concept being learned. Learning by analogy requires the ability to recognize the similarity between two problems and to use rules developed in one problem space in order to solve a problem in another problem space.

Likelihood

- [ISO GUIDE 73, 2009] Chance of something happening.
  - Note 1: In risk management terminology, the word likelihood is used to refer to the chance of something happening, whether defined, measured or determined objectively or subjectively, qualitatively or quantitatively, and described using general terms or mathematically [such as a probability (3.6.1.4) or a frequency (3.6.1.5) over a given time period].
  - Note 2: The English term likelihood does not have a direct equivalent in some languages; instead, the equivalent of the term probability is often used. However, in English, the probability is often narrowly interpreted as a mathematical term. Therefore, in risk management terminology, likelihood is used with the intent that it should have the same broad interpretation as the term probability has in many languages other than English.

Logic-based methods

- [Scharei et al., 2020] Logic-based methods consist of one out of several logics. They are often used in case of problem-solving.

Machine learning

- [Mamalet et al., 2021] Machine learning is a branch of artificial intelligence (AI) [ ] that refers to the automated detection of meaningful patterns in data. [It covers] a set of techniques that can learn from experience (input data).
- [Samoili et al., 2020] By learning, we refer to the ability of systems to automatically learn, decide, predict, adapt and react to changes, improving from experience, without being explicitly programmed. ML is widely included in the vast majority of efforts to identify AI categories, as the basic algorithmic approach to achieve AI regardless of the type of learning, namely reinforcement, supervised, semi-supervised, and unsupervised.
- [ISO/IEC 2382, 2015] Process by which a functional unit improves its performance by acquiring new knowledge or skills, or by reorganizing existing knowledge or skills.

Natural Language Processing (NLP)

- [Samoili et al., 2020] NLP, as the main task of communication, refers to the machines ability to identify, process, understand and/or generate information in written and spoken human communications. It is considered as an AI subdomain from several national strategies and AI experts, encompassing applications such as text generation, text mining, classification, and machine translation

Neural Network

- [ISO/IEC TR 24029-1, 2021] Network of primitive processing elements connected by weighted links with adjustable weights, in which each element produces a value by applying a non-linear function to its input values, and transmits it to other elements or presents it as an output value.

Pattern (artificial intelligence)

- [ISO/IEC 2382, 2015] Set of features and their relationships used to recognize an entity within a given context. These features may include a geometrical shape, a sound, a picture, a signal, or text.

Pattern Recognition

- [ISO/IEC 2382, 2015] Identification, by a functional unit, of physical or abstract patterns, and of structures and configurations.

Perception

- [Samoili et al., 2020] Perception refers to systems' ability to become aware of their environment through the senses: vision, hearing, and manipulation. etc., being vision and hearing the most developed areas in AI.

- [Scharei et al., 2020] Perception describes knowledge about the environment by transforming raw sensorial inputs into technically processable information. Sensorial inputs can be images, videos, or sound data.

Planning

- [Samoili et al., 2020] The main purpose of automated planning concerns the design and execution of strategies (e.g., an organised set of actions) to carry out some activity, and typically performed by intelligent agents, autonomous robots and unmanned vehicles. Unlike classical control and classification problems, the solutions are complex and must be discovered and optimised in the multidimensional space.
- [Scharei et al., 2020] Planning is the capability of setting and achieving goals. In more detail: a specific future state of the world is desirable, and the sequences of actions to access this state are highly relevant.

Planning (artificial intelligence)

- [ISO/IEC 2382, 2015] Process of deciding beforehand the manner and order of applying actions in order to reach a desired goal. Planning is performed with a view toward enhancing search efficiency and solving goal conflicts.
- [ISO/IEC DIS 22989, 2021a] Computational processes that compose a workflow out of a set of actions, aiming at reaching a specified goal.

Predictive Analytics

- [Logility, 2021] Prescriptive Analytics use optimization and simulation algorithms such as knowledge-based AI to advise on possible outcomes and answer: What should we do?

Preprocessed Data

- [Confiance.ai, 2021b] Raw Data (Filtered Raw Data or Corrected Raw Data) on which transformations have been applied to make data exploitable for a given use case.

Prescriptive Analytics

- [Logility, 2021] Predictive Analytics use statistical models and forecasting techniques such as ML to understand the future and answer: What could happen?

Production rule

- [ISO/IEC 2382, 2015] If-Then rule for representing knowledge in a rule-based system.

Reasoning (domain)

- [Samoili et al., 2020] The domain of reasoning tackles the way machines transform data into knowledge, or infer facts from data. Several classifications address knowledge representation and automated reasoning as a field of AI, to describe the process of justifying (reasoning) the available data and information, provide solutions and represent them efficiently, based on a set of symbolic rules

Recurrent neural networks (RNNs)

- [Techopedia, 2018a] A type of ANN that involves directed cycles in memory. One aspect of recurrent neural networks is the ability to build on earlier types of networks with fixed-size input vectors and output vectors, which make them well suited to operational applications, such as a self-driving car or self-flying airplane.

Reinforcement Learning

- [ISO/IEC TR 29119-11, 2020] Task of building a machine learning model using a process of trial and reward to achieve an objective. A reinforcement learning task can include the training of a ML model in a way similar to supervised learning plus training on unlabelled inputs gathered during the operation phase of the AI system. Each time the model makes a prediction, a reward is calculated, and further trials are run to optimize the reward.
- [EASA, 2021] This strategy is used in cases where there is an environment available for an agent to practise in. The agent(s) is(are) rewarded positively or negatively based on the effect of the actions on the environment. The ML model parameters are updated from this trial and-error sequence to optimise the outcome

Robustness

- [ISO/IEC DIS 22989, 2021b, ISO/IEC TR 24029-1, 2021] Ability of a system to maintain its level of performance under a variety of circumstances.
- [Mamalet et al., 2021] (Global) Ability of the system to perform the intended function in the presence of abnormal or unknown inputs / (Local) The extent to which the system provides equivalent responses for similar inputs.
- [EASA, 2021] For an input varying in a region of the state space, the system is producing the same outputs.
- [Gehr et al., 2018] Local robustness (or robustness, for short) requires that all samples in the neighbourhood of a given input are classified with the same label.

Scenario

- [ISO/PAS 21448, 2019] Description of the temporal development between several scenes in a sequence of scenes.

Scene

- [ISO/PAS 21448, 2019] Snapshot of the environment including the scenery, dynamic elements, and all actor and observer self representations, and the relationships between those entities.

Semantic network (semantic net)

- [ISO/IEC 2382, 2015] Concept-based knowledge representation in which objects or states appear as nodes connected with links that indicate the relationships between various nodes.

Statistical Bias

- [ISO/IEC TR 24027, 2021] Type of consistent numerical offset in an estimate relative to the true underlying value, inherent to most estimates

Supervised Learning

- [ISO/IEC 2382, 2015] Learning strategy in which the correctness of acquired knowledge is tested through feedback from an external knowledge source.
- [EASA, 2021] This strategy is used in cases where there is a labelled data set available to learn from. The ML algorithm processes the input data set, and a cost function measures the difference between the ML model output and the labelled data. The ML algorithm then adjusts the parameters to increase the accuracy of the ML model.

Symbolic methods

- [Scharei et al., 2020] Symbolic methods assume that the world can be reduced to symbol manipulations only. They can be subdivided into two different approaches: logic-based and knowledge-based.

Transparency

- [ISO/IEC DIS 22989, 2021b] <organization> Property of an organization that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner. <system> Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability.
- [Arrieta et al., 2020] A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models are divided into three categories: simulatable models, decomposable models and algorithmically transparent models.

- [Brundage et al., 2020] Making information about the characteristics of an AI devel opers operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.
- [IEEE 7000, 2021] Transparency means that information provided about a system is meaningful, useful, accessible, comprehensive, and truthful. A/ Meaningful means that information about a system should be relevant for users' concerns or user control. B/ Usefulness of information implies that consumers can act upon it and make choices easily, acting upon the information provided to them. C/ Accessible means that it is possible to easily obtain and retrieve the relevant information in a machine-readable or another way whether through state-of-the-art electronic channels or via constrained devices or constrained networks. D/ Comprehensive means that information about a system should be asy to read and understand for ordinary people and not require any expert knowledge. E/ Truthful means that information about a system accurately reflects a system's or system landscapes activities, such as data processing and data sharing practices. The information should be up to date and written in plain language that is clear and direct. It should not mislead users in any way, hide information, or give half-truths about practices.
  Related and Opposing values
  - Related values: Openness, cleanliness, explicability, explainability, access to data, auditability
  - Opposing values: Privacy, bribery, corruption
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability
- [ISO/IEC DIS 22989, 2021b] Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [Brundage et al., 2020] Making information about the characteristics of an AI developers operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.

Understandability

- [Arrieta et al., 2020] Understandability denotes the characteristic of a model to make a human understand its function Đ how the model works Đ without any need for explaining its internal structure or the algorithmic means by which the model processes data internally.

Unsupervised learning

- [EASA, 2021] This strategy is used in cases where there is no labelled data set available to learn from. The ML algorithm processes the data set, and a cost function indicates whether the ML model has converged to a stable solution. The ML algorithm then adjusts the parameters to increase the accuracy of the ML model.
- [ISO/IEC 2382, 2015] Learning strategy that consists in observing and analyzing different entities and determining that some of their subsets can be grouped into certain classes,

without any correctness test being performed on acquired knowledge through feedback from external knowledge sources. Once a concept is formed, it is given a name that may be used in subsequent learning of other concepts.

C.3. Data-driven AI

Adaptive neural network

- [ISO/IEC 2382, 2015] Neural network that is able to adjust its performance characteristics according to changes in its environment.

Adaptive learning

- [ISO/IEC 2382, 2015] Learning strategy that consists in adjusting internal knowledge according to advice from an external knowledge source, or transforming newly acquired information according to existing knowledge.

AI explainability

- [EASA, 2021] The AI explainability building block deals with the capability to provide the human with understandable and relevant information on how an AI/ML application is coming to its results.

AI safety risk

- [EASA, 2021] The AI safety risk mitigation building block considers that we may not always be able to open the AI black box to the extent required and that the safety risk may need to be addressed to deal with the inherent uncertainty of AI.

Analytic learning (explanation-based learning)

- [ISO/IEC 2382, 2015] Advanced form of deductive learning in which abstract or structured knowledge is derived from operational knowledge and from domain knowledge

Annotation Attribute

- [Confiance.ai, 2021b] Textual Name describing the information given by the Annotation.

Annotation Region

- [Confiance.ai, 2021b] Delimitation or sub-set of Xi Data on which Annotation will be performed. One Region may be associated with one or several Annotations.

Annotation Value

- [Confiance.ai, 2021b] Information Value (scalar, vector, category, textual É) associated with an Annotation Attribute.

Audio processing

- [Samoili et al., 2020] Audio processing refers to AI systems allowing the perception or generation (synthesis) of audio signals, including speech, but also other sound material (e.g. environmental sounds, music). Speech or voice recognition, audio processing or sound technologies are also often proposed to be archived as an AI subdivision.

Backward propagation

- [ISO/IEC TR 29119-11, 2020] Method used in artificial neural networks to determine the weights to be used on the network connections based on the computed error at the output of the network.

Bayesian network

- [ISO/IEC DIS 22989, 2021a] Probabilistic model that uses Bayesian inference for probability computations using a directed acyclic graph.

Bias

- [ISO/IEC TR 29119-11, 2020] Measure of the distance between the predicted value provided by the ML model and a desired fair prediction.
- [ISO/IEC TR 24028, 2020] Favouritism towards some things, people or groups over others.

Black box

- [ISO/IEC/IEEE 24765, 2017] 1. A system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or implementation are unknown or irrelevant. 2. Pertaining to an approach that treats a system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or irrelevant.

Classification

- [ISO/IEC TR 29119-11, 2020] Machine learning function that predicts the output class for a given input.

Classifier

- [ISO/IEC TR 29119-11, 2020] ML model used for classification.

Clustering

- [ISO/IEC TR 29119-11, 2020] Grouping of a set of objects such that objects in the same group (i.e. a cluster) are more similar to each other than to those in other clusters.

Completeness of explainability

- [Mamalet et al., 2021] Relates to the capability to describe a phenomenon in such a way that this description can be used to reach a given goal.

Component

- [Szyperski et al., 2002] A basic building-block for systems with well-defined interfaces, behavior and explicit context dependencies only.
  - A component can be deployed independently. That is, it implements a clear function.
  - A component can be composed with other components into systems, sub-system or new components.
  - A component can exist in the form of software or hardware or a combination of both.
- [RTCA DO-297, 2005] A self-contained hardware or software part, database, or combination thereof that may be configuration controlled.

Comprehensibility

- [Arrieta et al., 2020] When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. This notion of model comprehensibility stems from the postulates of Michalski, which stated that the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single chunks of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity.
- [Arrieta et al., 2020] when conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.

Computer vision

- [ISO/IEC DIS 22989, 2021a] Capability of a functional unit to acquire, process and interpret visual data.

Confusion matrix

- [ISO/IEC TR 29119-11, 2020] Table used to describe the performance of a classifier on a set of test data for which the true and false values are known.

Connectionism (connectionist paradigm)

- [ISO/IEC DIS 22989, 2021a] Form of cognitive modelling that uses a network of interconnected units which generally are simple computational units.

Convolutional neural networks (CNNs)

- [LeCun et al., 1989] A type of neural network for processing data that has a known grid-like topology. Convolutional networks use convolution, a specialised kind of linear operation in place of general matrix multiplication in at least one of their layers.

Data mining

- [ISO/IEC DIS 22989, 2021a] Computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing it, and summarizing potential relationships and impacts.

Data sciences

- [ISO/IEC 20546, 2019] Data science refers to the process for extracting knowledge from data Ð the approach can be either through exploration or by hypothesis testing. Data science refers to the complete data analytics lifecycle where data analytics is understood.
- [EASA, 2020] A broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the extraction of information and analysis to acquire knowledge from that information.

Data-driven AI

- [EASA, 2020] The data-driven approach focuses on building a system that can learn what is the appropriate answer based on having trained on a large number of examples.

Deep learning

- [ISO/IEC TR 29119-11, 2020] Approach to creating rich hierarchical representations through the training of neural networks with one or more hidden layers.
- [EASA, 2020] A specific type of machine learning based on the use of large (deep) neural networks to learn abstract representations of the input data through the use of multiple layers.

Descriptive Analytics

- [Logility, 2021] Descriptive Analytics use data aggregation and data mining to provide insight into the past and answer: What has happened?

Domain (distributed data processing)

- [ISO/IEC 2382, 2015] That part of a computer network in which the resources or addressing are under common control.

Explainability

- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.
- [Phillips et al., 2020] Explanation: Systems deliver accompanying evidence or reason(s) for all outputs. Meaningful: Systems provide explanations that are understandable to individual users. Explanation Accuracy: The explanation correctly reflects the system s process for generating the output. Knowledge Limits: The system only operates under the conditions for which it was designed or when the system reaches sufficient confidence in its output.
- [Arrieta et al., 2020] Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.
- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.

Explainable Artificial Intelligence (XAI)

- [Arrieta et al., 2020] An explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Explainable model

- [Phillips et al., 2020] An explanation of a model result is a description of how a model s outcome came to be.

Explanation facility

- [ISO/IEC 2382, 2015] Component of a knowledge-based system that explains how solutions were derived and justifies the steps used in reaching them.

F1-score (F-measure)

- [ISO/IEC TR 29119-11, 2020] Performance metric used to evaluate a classifier, which provides a balance (the harmonic average) between recall and precision.

False negative

- [ISO/IEC TR 29119-11, 2020] Incorrect reporting of a failure when in reality it is a pass

False positive

- [ISO/IEC TR 29119-11, 2020] Incorrect reporting of a pass when in reality it is a failure.

Feasible vs Infeasible Corner Case Data

- [SAE AS6983, 2019] Feasible Corner Case: Corner case that is part of the functional intent, thus inside the ML Model ODD

Feature engineering

- [ISO/IEC TR 29119-11, 2020] Activity in machine learning in which those attributes in the raw data that best represent the underlying relationships that should appear in the model are identified for use in the training data.
- [EASA, 2021] Feature engineering is a discipline consisting in transforming the pre-processed data so that it better represents the underlying structure of the data to be an input to the model training.

Generative adversarial networks (GANs)

- [Techopedia, 2018b] A type of construct in neural network technology that is composed of two neural networks: a generative network, that generates samples, and a discriminative network, that tries to detect whether a sample is real or the result of the generator.

Genetic algorithm

- [ISO/IEC DIS 22989, 2021a] Algorithm which simulates natural selection by creating and evolving a population of individuals (solutions) for optimization problems.

Global Robustness

- [Mamalet et al., 2021] Ability of the system to perform the intended function in the presence of abnormal or unknown inputs

Hyperparameter (machine learning)

- [Goodfellow et al., 2016] Settings that are used to control the behavior of the learning algorithm (e.g. number of hidden layers, learning rate, number of neurons per layer). The values of hyperparameters are not adapted by the learning algorithm.

Inference

- [ISO/IEC 2382, 2015] Reasoning by which conclusions are derived from known premises.

Information

- [ISO 9000, 2015] Meaningful data. [data being "facts about an object"]
- [Ackoff, 1989] Information represents the properties of objects and events, but in a more compact and useful form. The difference between data and information is functional, not structural. Information can be represented as descriptions, answers to questions that begin with such words as who, what, when, where, and how many.

Information analysis

- [ISO/IEC 2382, 2015] Systematic investigation of information and its flow in a real or planned system.

Inherent Data Quality

- [ISO/IEC 25012, 2008a] Degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.

Instantiation

- [ISO/IEC 2382, 2015] Substitution of a value for a variable, or creation of an example from a class. Example: A specific sick person is an instantiation of the generic object "patient". In a rule-based system, an instantiation is the result of successfully matching a rule against the contents of the knowledge base.

Interpretability

- [Arrieta et al., 2020] The ability to explain or to provide the meaning in understandable terms to a human.
- [Mamalet et al., 2021] Relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a human being. It is a basic requirement for an explanation.

Interpretable model

- An interpretable model should provide users with a description of what a stimulus, such as a datapoint or model output, means in context.

Learnability

- [ISO/IEC 25010, 2011a] Degree to which a product or system enables the user to learn how to use it with effectiveness, efficiency in emergency situations.

Learning

- [Scharei et al., 2020] Learning indicates incrementation of knowledge through experience.

Learning (machine learning)

- [ISO/IEC 2382, 2015] Process by which a biological or an automatic system gains knowledge or skills that it may use to improve its performance.
- [Samoili et al., 2020] By learning, we refer to the ability of systems to automatically learn, decide, predict, adapt and react to changes, improving from experience, without being explicitly programmed. ML is widely included in the vast majority of efforts to identify AI categories, as the basic algorithmic approach to achieve AI regardless the type of learning, namely reinforcement, supervised, semi-supervised, unsupervised.
- [ISO/IEC 2382, 2015] Process by which a functional unit improves its performance by acquiring new knowledge or skills, or by reorganizing existing knowledge or skills.
- [ISO/IEC TR 24028, 2020] Mathematical construct that generates an inference or prediction, based on input data.

Learning (neural networks)

- [ISO/IEC 2382, 2015] Process by which a neural network improves its performance by adjustment of its parameters in response to a succession of input patterns. In general, learning consists in connection weights adjustment.

Learning Assurance

- [EASA, 2021] The learning assurance building block is intended to cover the paradigm shift from programming to learning, as the existing development assurance methods are not adapted to cover learning processes specific to AI/ML.

Learning by analogy (associative learning)

- [ISO/IEC 2382, 2015] Learning strategy that combines inductive learning and deductive learning so that inductions determine the common characteristics of concepts being compared or associated, and deductions derive from these characteristics the features expected

of the concept being learned. Learning by analogy requires the ability to recognize the similarity between two problems and to use rules developed in one problem space in order to solve a problem in another problem space.

Likelihood

- [ISO GUIDE 73, 2009] Chance of something happening.
  - Note 1: In risk management terminology, the word likelihood is used to refer to the chance of something happening, whether defined, measured or determined objectively or subjectively, qualitatively or quantitatively, and described using general terms or mathematically [such as a probability (3.6.1.4) or a frequency (3.6.1.5) over a given time period].
  - Note 2: The English term likelihood does not have a direct equivalent in some languages; instead, the equivalent of the term probability is often used. However, in English, probability is often narrowly interpreted as a mathematical term. Therefore, in risk management terminology, likelihood is used with the intent that it should have the same broad interpretation as the term probability has in many languages other than English.

Local Robustness

- [Mamalet et al., 2021] The extent to which the system provides equivalent responses for similar inputs.

Machine learning

- [Mamalet et al., 2021] Machine learning is a branch of artificial intelligence (AI) which refers to the automated detection of meaningful patterns in data. [It covers] a set of techniques that can learn from experience (input data).

ML Model (resp. ML Constituent) ODD

- [SAE AS6983, 2019] The ODD is defined as all the foreseeable operating conditions under which a ML Model (resp. ML Constituent) is expected to work. This may be the system operating domain, a superset of it to provide robustness, or a subset of it to limit the design to a feasible region.

ML Robustness

- [SAE AS6983, 2019] The capacity of an ML model to preserve its expected / intended performance under well-characterized abnormalities or deviations to its inputs and operating conditions outside its operational design domain (ODD)

ML Stability

- [SAE AS6983, 2019] The capacity of an ML model to preserve its expected / intended performance under well-characterized and bounded perturbations to its inputs and operating conditions within its operational design domain (ODD)

Natural Language Processing (NLP)

- [Samoili et al., 2020] NLP, as the main task of communication, refers to the machine s ability to identify, process, understand and/or generate information in written and spoken human communications. It is considered as an AI subdomain from several national strategies and AI experts, encompassing applications such as text generation, text mining, classification, and machine translation

Neural Network

- [ISO/IEC TR 24029-1, 2021] Network of primitive processing elements connected by weighted links with adjustable weights, in which each element produces a value by applying a non-linear function to its input values, and transmits it to other elements or presents it as an output value.

Operational Design Domain (ODD)

- [SAE J3016, 2018] Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristic.

Outlier

- [ISO/DIS 5725-1, 2020] A value from a set of values that is inconsistent with the other values of that set, identified by a statistical test.
- [ISO 16269-4, 2010] Member of a small subset of observations that appears to be inconsistent with the remainder of a given sample.
  - Note 1: The classification of an observation or a subset of observations as outlier(s) is relative to the chosen model for the population from which the data set originates. This or these observations are not to be considered as genuine members of the main population.
  - Note 2: An outlier may originate from a different underlying population, or be the result of incorrect recording or gross measurement error.
  - Note 3: The subset may contain one or more observations.

Outlier Data

- [Holloway, 2019] An inlier is a data value that lies in the interior of the ODD following an error during data management. A simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius. Because inliers are difficult to distinguish from good data values they are sometimes difficult to find and correct.

Output Data (of a module)

- [Confiance.ai, 2021b] Output data are data obtained as the output of a given module of a machine learning workflow no matter what the stage in the workflow. Inlier, Outlier, Novelty, or Infeasible Corner Case Data

Overfitting

- [ISO/IEC TR 29119-11, 2020] Generation of a machine learning model that corresponds too closely to the training data, resulting in a model that finds it difficult to generalize to new data.

Pattern (artificial intelligence)

- [ISO/IEC 2382, 2015] Set of features and their relationships used to recognize an entity within a given context. These features may include a geometrical shape, a sound, a picture, a signal, or text.

Pattern Recognition

- [ISO/IEC 2382, 2015] Identification, by a functional unit, of physical or abstract patterns, and of structures and configurations.

Perception

- [Samoili et al., 2020] Perception refers to systems ability to become aware of their environment through the senses: vision, hearing, manipulation. etc., being vision and hearing the most developed areas in AI.
- [Scharei et al., 2020] Perception describes knowledge about the environment by transforming raw sensorial inputs into technically processable information. Sensorial inputs can be images, videos, or sound data.

Precision of Explanability

- [Mamalet et al., 2021] Indicates how much details must be provided to the human to let her/him execute mentally the inference in a right way with respect to her/his goal. For instance, there is no need to know the laws of general relativity or quantum mechanics to predict the trajectory of a ball.

Prediction Dataset

- [Confiance.ai, 2021b] This set must be identified because a comparison between Model Predictions and Ground Truth Observations will be performed to characterize model performance during the ML Model development.

Predictive Analytics

- [Logility, 2021] Prescriptive Analytics use optimization and simulation algorithms such as knowledge based AI to advise on possible outcomes and answer: What should we do?
- [Logility, 2021] Predictive Analytics use statistical models and forecasting techniques such as ML to understand the future and answer: What could happen?

Probabilistic methods

- [Scharei et al., 2020] Probabilistic methods allow acting in incomplete information scenarios.

Recurrent neural networks (RNNs)

- [Techopedia, 2018a] A type of ANN that involves directed cycles in memory. One aspect of recurrent neural networks is the ability to build on earlier types of networks with fixed-size input vectors and output vectors, which make them well suited to operational applications, such as a self-driving car or self-flying airplane.

Reinforcement Learning

- [ISO/IEC TR 29119-11, 2020] Task of building a machine learning model using a process of trial and reward to achieve an objective. A reinforcement learning task can include the training of a ML model in a way similar to supervised learning plus training on unlabelled inputs gathered during the operation phase of the AI system. Each time the model makes a prediction, a reward is calculated, and further trials are run to optimize the reward.
- [EASA, 2021] This strategy is used in cases where there is an environment available for an agent to practice in. The agent(s) is(are) rewarded positively or negatively based on the effect of the actions on the environment. The ML model parameters are updated from this trial and-error sequence to optimise the outcome

Statistical Bias

- [ISO/IEC TR 24027, 2021] Type of consistent numerical offset in an estimate relative to the true underlying value, inherent to most estimates

Supervised Learning

- [ISO/IEC 2382, 2015] Learning strategy in which the correctness of acquired knowledge is tested through feedback from an external knowledge source.

- [EASA, 2021] This strategy is used in cases where there is a labelled data set available to learn from. The ML algorithm processes the input data set, and a cost function measures the difference between the ML model output and the labelled data. The ML algorithm then adjusts the parameters to increase the accuracy of the ML model.

Syntactic Data Accuracy

- [ISO/IEC 25012, 2008a] Closeness of the data values to a set of values defined in a domain considered syntactically correct.

Test Dataset

- [Confiance.ai, 2021b] A Dataset that is only composed of Observations that will be used only for evaluating the ML Model performance under operational configuration. Test Dataset must be fully independent of Training Dataset and Validation Dataset. No Observation from this set can be part of these two Datasets.

Training Dataset

- [Confiance.ai, 2021b] A Dataset that is only composed of Observations that will be used only for the training of the ML Model.

Trueness

- [ISO/DIS 5725-1, 2020] Closeness of agreement between the expectation of test results and a true value Note 1: The measure of trueness is usually expressed in terms of bias. Note 2: Trueness is sometimes referred to as accuracy of the mean. This usage is not recommended. Note 3: In practice, the accepted reference value is substituted for the true value.

Unsupervised learning

- [EASA, 2021] This strategy is used in cases where there is no labelled data set available to learn from. The ML algorithm processes the data set, and a cost function indicates whether the ML model has converged to a stable solution. The ML algorithm then adjusts the parameters to increase the accuracy of the ML model.
- [ISO/IEC 2382, 2015] Learning strategy that consists in observing and analyzing different entities and determining that some of their subsets can be grouped into certain classes, without any correctness test being performed on acquired knowledge through feedback from external knowledge sources. Once a concept is formed, it is given a name that may be used in subsequent learning of other concepts.

Validation Dataset

- [Confiance.ai, 2021b] A Dataset that is only composed Observations that will be used only for evaluating the generalization capabilities of the ML Model and choosing / tuning Hyperparameters values. In that context, Validation Dataset is composed of Observations that are not already in the Training Dataset.

C.4. Knowledge-based AI

Agent

- [ISO/IEC DIS 22989, 2021a] Automated entity that perceives its environment and takes actions to achieve its goals.

Domain (artificial intelligence)

- [ISO/IEC 2382, 2015] Specific field of knowledge or expertise.

Domain knowledge

- [ISO/IEC 2382, 2015] Knowledge accumulated in a particular domain.

Domain model

- [ISO/IEC 2382, 2015] Model of a specific field of knowledge or expertise.

Expert system

- [ISO/IEC 2382, 2015] Computer system that provides for expertly solving problems in a given field or application area by drawing inferences from a knowledge base developed from human expertise. The term "expert system" is sometimes used synonymously with "knowledge-based system", though it is usually taken to emphasize expert knowledge. Some expert systems are able to improve their knowledge base and develop new inference rules based on their experience with previous problems.

Heuristic method

- [ISO/IEC 2382, 2015] Any exploratory method of solving problems in which an evaluation is made of the progress towards an acceptable final result using a series of approximate results, for example by a process of guided trial and error.

Heuristic rule

- [ISO/IEC 2382, 2015] Ad hoc rule written to formalize the knowledge and experience an expert uses to solve a problem.

Heuristic Search

- [ISO/IEC 2382, 2015] Search, based on experience and judgment, used to obtain acceptable results without guarantee of success.

Hierarchical Planning

- [ISO/IEC 2382, 2015] Planning that refines the vague parts of a plan into more detailed subplans by generating a hierarchical representation of it.

Inference

- [ISO/IEC 2382, 2015] Reasoning by which conclusions are derived from known premises.

Inference engine

- [ISO/IEC 2382, 2015] Component of an expert system that applies principles of reasoning to draw conclusions from representations of information stored in a knowledge base.

Instantiation

- [ISO/IEC 2382, 2015] Substitution of a value for a variable, or creation of an example from a class.
  Example: A specific sick person is an instantiation of the generic object "patient".
- In a rule-based system, an instantiation is the result of successfully matching a rule against the contents of the knowledge base.

Logic-based methods

- [Scharei et al., 2020] Logic-based methods consist of one out of several logics. They are often used in case of problem-solving.

Natural Language Processing (NLP)

- [Samoili et al., 2020] NLP, as the main task of communication, refers to the machineÕs ability to identify, process, understand and/or generate information in written and spoken human communications. It is considered as an AI subdomain from several national strategies and AI experts, encompassing applications such as text generation, text mining, classification, and machine translation

Planning

- [Samoili et al., 2020] The main purpose of automated planning concerns the design and execution of strategies (e.g., an organized set of actions) to carry out some activity, and typically performed by intelligent agents, autonomous robots, and unmanned vehicles.

Unlike classical control and classification problems, the solutions are complex and must be discovered and optimized in the multidimensional space.
- [Scharei et al., 2020] Planning is the capability of setting and achieving goals. In more detail: a specific future state of the world is desirable, and the sequences of actions to access this state are highly relevant.

Planning (artificial intelligence)

- [ISO/IEC 2382, 2015] Process of deciding beforehand the manner and order of applying actions in order to reach the desired goal. Planning is performed with a view toward enhancing search efficiency and solving goal conflicts.
- [ISO/IEC DIS 22989, 2021a] Computational processes that compose a workflow out of a set of actions, aiming at reaching a specified goal.

Prescriptive Analytics

- [Logility, 2021] Predictive Analytics use statistical models and forecasting techniques such as ML to understand the future and answer: What could happen?

Production rule

- [ISO/IEC 2382, 2015] If-Then rule for representing knowledge in a rule-based system.

Reasoning (domain)

- [Samoili et al., 2020] The domain of reasoning tackles the way machines transform data into knowledge, or infer facts from data. Several classifications address knowledge representation and automated reasoning as a field of AI, to describe the process of justifying (reasoning) the available data and information, provide solutions and represent them efficiently, based on a set of symbolic rules

Semantic network (semantic net)

- [ISO/IEC 2382, 2015] Concept-based knowledge representation in which objects or states appear as nodes connected with links that indicate the relationships between various nodes.

Symbolic methods

- [Scharei et al., 2020] Symbolic methods assume that the world can be reduced to symbol manipulations only. They can be subdivided into two different approaches: logic-based and knowledge-based.

Chapter D

Data Engineering

D.1. Introduction

Characterizing datasets and assessing data quality is essential at the beginning of a project, as well as all along the project lifecycle. It is a prerequisite to be able to develop performant and trustable data-driven applications and models. In this context, research on data quality assessment has been conducted in numerous fields (see state of the art in [Gudivada et al., 2017] and has evolved from studies related to traditional enterprise and software information systems [Batini et al., 2009, Pipino et al., 2002], to studies related to the challenges of data quality for big data applications [Cai and Zhu, 2015] and machine learning applications [Gudivada et al., 2017, Picard et al., 2020], and also [Confiance.ai, 2021c, Data exploration and qualification]).

Data quality plays a significant role in machine learning applications because model outputs are conditioned by the representativity of the inputs. All the dimensions of data quality related to reliability (such as accuracy, completeness, consistency), availability, relevance, and usability [Batini et al., 2009, Cai and Zhu, 2015] are key components to ensure that inputs are representative. To illustrate the problems encountered when datasets are not qualified, on one hand, a "computationally large" dataset can still be a "statistically sparse" dataset as stated in [Gudivada et al., 2017]. Hence it can create some quality problems in the results when the chosen algorithm falls in the sparse regions of the dataset. On the other hand, even when a dataset is qualified and evaluated as 98% correct, still the two remaining percent can generate problems and unexpected results. No matter what the model goal or the type of input data, all the projects share the same needs for data characterization before starting any models, as it will guide the choice of algorithms/architecture, and helps to understand and analyze the model's performance later on. To this end, different methodologies can be adopted to assess data quality [Batini et al., 2009]. All those methodologies rely on the same strategy based on three main steps:

- *State reconstruction*, where contextual information is collected on the data itself and the

data acquisition process.
- *Assessment/measurement*, where the dataset is analyzed through different measures/dimensions to get a view of the data quality.
- *Improvement*, where different techniques are deployed (data correction, new data collection, etc.) to meet data quality targets.

The implementation of those steps will depend on the type of data and the goal of the application. Qualitative methods for example based on data visualization, or quantitative methods using metrics to minimize or maximize can be used to describe and evaluate datasets based on definitions, requirements, and verification plans [Picard et al., 2020] that will allow monitoring the dataset characteristics and qualities through the project lifecycle.

D.2. Data Engineering Taxonomy

Anomaly

- [SAE AS6983, 2019] Data which is outside the ML Model ODD

Anomaly - misclassified samples

- [Corbière et al., 2019, Geifman and El-Yaniv, 2017] Objects that are likely to be misclassified and that fall near the decision boundary where the classifier is uncertain. Such problems are known as the problem of classification with reject option.

Anomaly - novelty detection

- [Schölkopf et al., 2001] Test points that could be new observations, i.e., the equivalent of an outlier for the test data. An example would be a new rare breed of dogs for a dataset of dog breeds. Such points are particularly interesting in active learning where once identified they are annotated and added to the training set, further improving the current classifier.

Anomaly - outlier detection

- [Rousseeuw and Driessen, 1999] Data points from the training set that are far from the others, i.e., an unusual or noisy training sample. An example would be a highly blurry picture of a pedestrian in a dataset for pedestrian detection or a picture of cat within a dataset of dogs.

Anomaly - out-of-distribution (OOD) detection

- [Liang et al., 2017, Malinin and Gales, 2018, Meinke and Hein, 2019] Objects that are drawn from a distribution different from the training distribution. In deep learning, we can distinguish two types of OOD:

  - low-level: different pixel statistics due to a mismatch between training and testing environments, e.g., a perception model for a vehicle trained for a country and tested on another one, a model trained on day-time images and tested on night-time images.
  - high-level (semantic): unknown objects or entities in a familiar environment, e.g., electric scooters for a perception model trained before scooters populated the streets

Augmented Data

- [Confiance.ai, 2021b, Data Taxonomy] Augmented data are data generated via the modification of existing Observations or by the creation of new synthetic Observations in order to obtain a Dataset covering more situations expected in the operational domain.

Bias

- [ISO/IEC TR 29119-11, 2020] Measure of the distance between the predicted value provided by the ML model and a desired fair prediction.
- [ISO/IEC TR 24028, 2020] Favouritism towards some things, people or groups over others.

Black box

- [ISO/IEC/IEEE 24765, 2017] 1. A system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or implementation are unknown or irrelevant. 2. Pertaining to an approach that treats a system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or irrelevant.

Confidentiality

- [ISO/IEC 25010, 2011a] Degree to which the prototype ensures that data are accessible only to those authorized to have access.

Contextual Data

- [Confiance.ai, 2021b, Data Taxonomy] This data is composed of Annotations and all information that are not Observations (X_i Data, Observed Y_i Data) used to train or evaluate the ML Model. For example, they are used as criteria to search Data, inside the Database, as explicative variables in the context of model performance analysis. A Contextual Data is linked to a X_i Data, a Y_i Data, an Observation, a Prediction, a Data Selection, a Dataset, or a Database.

Corner Case Data

- Where an edge case involves pushing one ML Model ODD parameter to a minimum or maximum, putting the data at the "edge" of the ML Model ODD, a corner case involves

doing so with 2 or more ML Model ODD parameters, which would put data at a "corner" of a multidimensional ML Model ODD [adapted from https://arxiv.org/abs/2103.03678]

Correct data

- [SAE AS6983, 2019] A nominal data is a data value that lies in the interior of the ML Model ODD statistical distribution and is correct with respect to the ML requirements

Corrected Raw Data

- [Confiance.ai, 2021b, Data Taxonomy] Raw Data on which modifications have been applied once and for all to correct an error or make data exploitable for a given use case.

Data

- [ISO/IEC 2382, 2015] Re-interpretable representation of information in a formalized manner suitable for communication, interpretation or processing. Data can be processed by human or automatic means.
- [Ackoff, 1989] Data are provided to the intelligent reasoning engines in order to give them a representation of the world. Once the data samples are acquired by the systems, they become information. This information is then interpreted by the intelligent systems, according to the context related to the acquisition of the data items, and the domain knowledge, among others. The domain knowledge is used to interpret a set of data items within a specific context.

Data Accessibility

- [ISO/IEC 25012, 2008a] Degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.

Data Accessibility

- [Mamalet et al., 2021] The effort required to access data

Data Accuracy

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.
- [Mamalet et al., 2021] Accuracy depends on data gathering/generation and measures the faithfulness to the real value. It also measures the degree of ambiguity of the representation of the information.

Data annotation

- [Confiance.ai, 2021b, Data Taxonomy] Annotation is the manual or (partially) automated process which aims at adding descriptive information to X_i Data, Y_i Data, Observations, Predictions, data Selection, Dataset, or Database.
- [Confiance.ai, 2021b, Data Taxonomy] Annotation is the descriptive information itself that is output by the Annotation process above. This descriptive information may have very different data types: textual, categorial, numerical, image region associated with a category É Again, this Annotation is linked, or associated, to X_i Data, Y_i Data, Observations, Predictions, data Selection, Dataset or Database. In the context of an Observation (X_i, Observed Y_i) part or all of X_i and Y_i may be composed of Annotations.
- [ISO/IEC DIS 22989, 2021a] Process of attaching a set of descriptive information to data without any change to that data.

Data augmentation

- [ISO/IEC DIS 22989, 2021a] Process of creating new data samples by manipulating the original data.

Data availability

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.

Data characterizing

- [Confiance.ai, 2021b, Data Taxonomy] Transformation, augmentation, and annotation in the purpose of data preparation and featuring engineering for the design of a model

Data completeness

- [ISO/IEC 25012, 2008a] Degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
- [EASA, 2020] A dataset is considered to be complete if it is comprehensive in the sense that it has been sampled properly to cover the specified space of the ODD (Operational Design Domain) of the intended application.
- [ED-76A, 2015] The degree of confidence that all of the data needed to support the intended use is provided

Data compliance

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.

Data confidentiality

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.

Data consistency

- [ISO/IEC 25012, 2008a] 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.
- [Mamalet et al., 2021] The deviation of values, domains, and formats between the original dataset and a pre-processed dataset

Data correctness

- [ED-76A, 2015] Data meeting stated quality requirements.

Data currency

- Refers to the degree to which data has attributes that are of the right age in a specific context of use.
- It is the degree to which a datum is up to date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value.

Data diversity

- [Ashmore and Madahar, 2019] Data diversity is achieved by using different data sets, since those sets should capture the essence of requirements and any drawback impacts their fulfilment by the respective ML modules. The requirements can be achieved globally, i.e. relying upon different data sources, or locally, i.e., relying upon different subsets of the same source.

Data efficiency

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.

Data engineering

- [Confiance.ai, 2021c] The discipline that aims to organize, structure, trace and select data in such a way that its quality, availability, relevance and traceability can be guaranteed throughout the life cycle of the data.
- [ISO/IEC 20546, 2019] Discipline which is related to the engineering aspects of systems, processing, models and management of data, including but not limited to big data.

Data governance

- [EASA, 2021] The capability of an organization to ensure that high data quality exists throughout the complete life cycle of the data, and data controls are implemented that support business objectives. The key focus areas of data governance include data availability, usability, consistency, integrity, and sharing.

Data integrity

- [ED-76A, 2015] A degree of assurance that aeronautical data and its value has not been lost or altered (since the data origination or authorized amendment)
- [ISO/IEC TR 29119-11, 2020] Property whereby data have not been altered in an unauthorized manner since they were created, transmitted, or stored.

Data management

- [EASA, 2021] The data management process will capture the requirements to be transferred to the implementation, regarding the pre-processing and feature engineering to be performed on the inference model.

Data mining

- [ISO/IEC DIS 22989, 2021a] Computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing it, and summarizing potential relationships and impacts.

Data portability

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.

Data precision

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that are exact or that provide discrimination in a specific context of use.

Data privacy

- [ISO 17572, 2015] Rights and obligations of individuals and organizations with respect to the collection, use, retention, disclosure and disposal of personal information
- Data privacy is the protection of personal data from those who should not have access to it and the ability of individuals to determine who can access their personal information.

Data processing

- [Confiance.ai, 2021b, Data Taxonomy] Acquisition and transformation in the purpose of cleaning and filtering raw data

Data quality

- [ISO/IEC 25024, 2015] Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions.
- [Mamalet et al., 2021] The extent to which data are free of defects and possess desired features
- [ISO/IEC 25012, 2008a] The grounds where the system for assessing the quality of data products is built on.

Data quality measure

- [ISO/IEC 25024, 2015] Variable to which a value is assigned as the result of measurement of a data quality characteristic

Data quality model

- [ISO/IEC 25024, 2015] Defined set of characteristics which provides a framework for specifying data quality requirements and evaluating data quality

Data recoverability

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.

Data reliability

- For the uses intended, subject (data) elements that demonstrate accuracy, completeness, integrity, stability, repeatability and precision.
- [EASA, 2021] Confidence level in the goodness of the data (e.g. because it is provided by a trusted source, or by a high-fidelity model).

Data representativeness

- [Mamalet et al., 2021] Refers in statistics to the notion of sample and population. Transposed to AI, the sample corresponds to the dataset available for the development of the model (training, validation, testing), and the population corresponds to all possible.
- [EASA, 2021] A dataset is representative when it is complete, and the distribution of its key characteristics is similar to the intended space of the ODD of the targeted application. Representativeness includes completeness.

Data sampling

- [ISO/IEC DIS 22989, 2021a] Process to select a subset of data samples intended to present patterns and trends similar to that of the larger dataset being analysed.

Data sciences

- [ISO/IEC 20546, 2019] Data science refers to the process for extracting knowledge from data Ð the approach can be either through exploration or by hypothesis testing. Data science refers to the complete data analytics lifecycle where data analytics is understood.
- [EASA, 2020] A broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the extraction of information and analysis to acquire knowledge from that information.

Data Scoping

- [Confiance.ai, 2021b, Data Taxonomy] Selection and allocation of the different observations and associated features in the different training, testing, and validating set

Data Security

- [ISO/IEC 27000, 2018] Preservation of confidentiality, integrity and availability of data. In addition, other properties, such as authenticity, accountability, non-repudiation, and reliability can also be involved.

Data Selection

- [Confiance.ai, 2021b, Data Taxonomy] Any set of data that needs to be identified specifically, whatever the type of Data (Xi Data, Y_i Data, Observations, Predictions, Contextual É). A data selection can be defined for all manipulation tasks that will be performed equally on all the data in the selection and that may require traceability or persistency.

Data Timeliness

- [Mamalet et al., 2021] The time delay from data generation and acquisition to utilization. If required data cannot be collected in real time or if the data need to be accessible over a very long time and are not regularly updated, then information can be outdated or invalid.
- [EASA, 2021] This property ensures that data are up-to-date and not obsolete.

Data Traceability

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.
- [Mamalet et al., 2021] Reflects how much both the data source and the data pipeline are available. Activities to identify all the data pipeline components have to be considered in order to guarantee such quality.

Data Understandability

- [ISO/IEC 25012, 2008a] Degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.

Data Usability

- [Mamalet et al., 2021] Quality bound to the credibility of data, i.e. if their correctness is regularly evaluated, and if data exist in the range of known or acceptable values.

Database

- [Confiance.ai, 2021b, Data Taxonomy] Set of all available Data, whatever they are used or not during the ML lifecycle and whatever the type of Data (X_i Data, Y_i Data, Observations, Predictions, Contextual)

Dataset

- [Branco et al., 2008] An aggregation of data, typically spawning more than one physical file,that are processed together and serve collectively as input or output of a computation or data ac-quisition process.
- [Confiance.ai, 2021b, Data Taxonomy] A set of Data that is only composed of (Xi, Yi) for i=1...N Observations or Predictions with no possible mixing between both types. Its goal is to be used to train a model or evaluate its performance. As Contextual Data are not used by the model, they cannot be part of this set.

Domain (distributed data processing)

- [ISO/IEC 2382, 2015] That part of a computer network in which the resources or addressing are under common control.

Edge Case Data

- Situation or data that occurs only at an extreme (maximum or minimum) operating parameter within the ML Model ODD [adapted from https://arxiv.org/abs/2103.03678]

Elementary Data

- [Confiance.ai, 2021b, Data Taxonomy] The finest level of data that can be manipulated for a specific Use Case. It is a subpart of X_i Data or Y_i Data that will not be used as such directly for training but that may be manipulated in pre-processing or post-processing stages of the ML Life Cy

Extended Dataset

- [Confiance.ai, 2021b, Data Taxonomy] A set of Data that is composed of (Xi, Yi) for i=1...N Observations or Predictions (with no possible mixing between both types) and Contextual Data associated with those Observations / Predictions

Extended Dataset

- [Confiance.ai, 2021b, Data Taxonomy] A set of Data that is composed of (Xi, Yi) for i=1...N Observations or Predictions (with no possible mixing between both types) and Contextual Data associated with those Observations / Predictions

Feasiblle vs Infeasible Corner Case Data

- [SAE AS6983, 2019] Feasible Corner Case: Corner case that is part of the functional intent, thus inside the ML Model ODD

Feature engineering

- [ISO/IEC TR 29119-11, 2020] Activity in machine learning in which those attributes in the raw data that best represent the underlying relationships that should appear in the model are identified for use in the training data.

Feature engineering

- [EASA, 2021] Feature engineering is a discipline consisting in transforming the pre-processed data so that it better represents the underlying structure of the data to be an input to the model training.

Filtered Raw Data

- [Confiance.ai, 2021b, Data Taxonomy] Raw Data on which filters have been applied: filtered raw data are not transformed, they are only filtered through selection processes.

Goal Structuring Notation

- [Group et al., 2018] Graphical argumentation notation that can be used to document explicitly the individual elements of any argument (claims, evidence and contextual information) and, perhaps more significantly, the relationships that exist between these elements

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Inherent Data Quality

- [ISO/IEC 25012, 2008a] Degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.

Inlier Data

- Data which is within the ML Model ODD according to the existing ML Model ODD parameters, but which should have been considered outside the ML Model ODD if it had been correctly described with the introduction of at least one new ML Model ODD parameter. A novelty is in general due to a lack of characterization of the ML Model ODD. It could be integrated to the ML Model ODD after analysis following the upgrade policy of the ML Model ODD. A novelty that is already outside the ML Model ODD is therefore an outlier. [Insprired from Glossary of statistical terms
, https://stats.oecd.org/glossary/detail.asp?ID=3464]

Input Data (of a module)

- [Confiance.ai, 2021b, Data Taxonomy] Input data are data passed into a given module of a machine learning workflow no matter what the stage in the workflow.

Label (organisation of data)

- [ISO/IEC 2382, 2015] Identifier that is attached to a set of data elements.

Likelihood

- [ISO GUIDE 73, 2009] Chance of something happening.
  - Note 1: In risk management terminology, the word likelihood is used to refer to the chance of something happening, whether defined, measured or determined objectively or subjectively, qualitatively or quantitatively, and described using general terms or mathematically [such as a probability or a frequency over a given time period].
  - Note 2: The English term likelihood does not have a direct equivalent in some languages; instead, the equivalent of the term probability is often used. However, in English, the probability is often narrowly interpreted as a mathematical term. Therefore, in risk management terminology, likelihood is used with the intent that it should have the same broad interpretation as the term probability has in many languages other than English.

Localized Annotation

- [Confiance.ai, 2021b, Data Taxonomy] An Annotation that is composed of a Region, an Annotation Attribute, and an Annotation Value.

Metadata

- [Confiance.ai, 2021b, Data Taxonomy] Metadata corresponds to Contextual Data associated with a full Data Selection, a Dataset, or a Database. Thus, Metadata is a particular case of Contextual Data.

Novel Data = Novelty

- [SAE AS6983, 2019] Infeasible Corner Case: Corner case that is not part of the functional intent, thus outside the ML Model ODD

Object and Event Detection and Response

- [NHTSA et al., 2016] Detection by the driver or Higly Automated Vehicle (HAV) system of any circumstance that is relevant to the immediate driving task, as well as the implementation of the appropriate driver or HAV system response to such circumstance. For purposes of this Guidance, the HAV system is responsible for performing the OEDR while in its ODD and automation is engaged.

Output Data (of a module)

- [Confiance.ai, 2021b, Data Taxonomy] Output data are data obtained as the output of a given module of a machine learning workflow no matter what the stage in the workflow. Inlier, Outlier, Novelty, or Infeasible Corner Case Data

Preprocessed Data

- [Confiance.ai, 2021b, Data Taxonomy] Raw Data (Filtered Raw Data or Corrected Raw Data) on which transformations have been applied to make data exploitable for a given use case.

Raw Data

- [Confiance.ai, 2021b, Data Taxonomy] Data or information directly coming from the sensor or more generally from the data acquisition process. We refer to raw data even when the data acquisition process is configurable (e.g. color filter, white balance, etc.) if it is not possible to change a posterior the data coming from the sensor without the application of an external transformation.

Semantic Data Accuracy

- [ISO/IEC 25012, 2008a] Closeness of the data values to a set of values defined in a domain considered semantically correct.

Simple Annotation (Tag)

- [Confiance.ai, 2021b, Data Taxonomy] An Annotation that is composed of an Annotation Attribute and an Annotation Value.

Statistical Bias

- [ISO/IEC TR 24027, 2021] Type of consistent numerical offset in an estimate relative to the true underlying value, inherent to most estimates

Syntactic Data Accuracy

- [ISO/IEC 25012, 2008a] Closeness of the data values to a set of values defined in a domain considered syntactically correct.

Test Dataset

- [Confiance.ai, 2021b, Data Taxonomy] A Dataset that is only composed of Observations that will be used only for evaluating the ML Model performance under operational configuration. Test Dataset must be fully independent of Training Dataset and Validation Dataset. No Observation from this set can be part of these two Datasets.

Training Dataset

- [Confiance.ai, 2021b, Data Taxonomy] A Dataset that is only composed of Observations that will be used only for the training of the ML Model.
- [Confiance.ai, 2021b, Data Taxonomy] A Dataset that is only composed of Observations that will be used only for the training of the ML Model.

Validation Dataset

- [Confiance.ai, 2021b, Data Taxonomy] A Dataset that is only composed Observations that will be used only for evaluating the generalization capabilities of the ML Model and choosing / tuning Hyperparameters values. In that context, Validation Dataset is composed of Observations that are not already in the Training Dataset.

Chapter E

Knowledge Engineering

E.1. Introduction

In its initial form, Knowledge Engineering (KE) focused on the transfer process; transferring the expertise of a problem-solving human into a program that could take the same data and make the same conclusions. Thus, KE dealt with the development of information systems in which knowledge and reasoning play pivotal roles.

In the 1990s, the attention of the KE community shifted gradually to domain knowledge, in particular reusable representations in the form of ontologies[1].

This evolution aimed at alleviating KE limitation to accurately reflect how humans make decisions and more specifically its failure to take into account intuition and "gut feeling", known as "reasoning by analogy". Today, KE refers to the process of understanding and then representing human knowledge in data structures, semantic models (conceptual diagram of the data as it relates to the real world), and heuristics (rules in AI context). KE uses a modeling process that creates a system that gives the same results as the expert without following the same path or using the same information sources.

Moreover, to support decision making, the situation consists of the context and the preferences of the decision-maker, then, we have to take into account that:

- A decision is goal-oriented, and often there are multiple, conflicting goals to be met by the same decision.
- Multi objectives or multi criteria decision could be done.
- Quality of decision is important.

[1] In computer science and information science, an ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many, or all domains of interest. More simply, an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.

- Explanation is expected by decision-makers and/or stakeholders

KE goal it to acquire the knowledge related to a specific task from expert people, and to transform it into *if-then rules* in order to make deductions or choices. The basic assumption is that both knowledge and experience can be captured and archived in textual or rule-based form, using formalization methods.

Knowledge Representation and Reasoning (KRR) is a key area of knowledge-based AI that focuses on designing computer representations that capture information and available knowledge about the world in order to solve complex problems or to understand human knowledge in data structures, semantic models (conceptual diagram of the data as it relates to the real world) and heuristics (rules leading to solutions in an AI approach).

KRR is responsible for representing information about the real world so that a computer can understand and can utilize this knowledge to solve complex real world problems such as diagnosis a medical condition or communicating with humans in natural language. It is also a way which describes how we can represent knowledge in artificial intelligence. Knowledge representation is not just storing data into some database, but it also enables an intelligent machine to learn from that knowledge and experiences so that it can behave intelligently like a human. In KRR, a fundamental assumption is that knowledge is explicitly represented in a declarative form, suitable for processing by dedicated reasoning engines to be useful and helpful in achieving certain tasks. The same information may be represented in many different ways, depending on how you want to use that information.

Disciplines such as statistics, decision theory and operations research developed various methods for making rational decisions. Such methods, enhanced by techniques coming from information science and AI, have been implemented, either as stand-alone tools or as integrated computing environments for complex decision making. Simultaneously, the new standards in knowledge representation (W3C) now allows us to consider knowledge both as structured data that can be processed by machine learning methods and as symbolic knowledge data that can be processed by inference engines.

Moreover, knowledge representation and reasoning are key:

- Decision and problem solving could be under certainty, under risk, under uncertainty;
- Completeness, Soundness of knowledge are important issues.

And last but not least, constraints are limitations imposed by context or by resources that do not allow stakeholders or the decision support system to take certain actions. That involves: discrete (e.g. resources); continuous (e.g. distance constraints) and operational constraints (e.g. real-time).

By essence, a knowledge-based system (KBS) aims to assist rather than replace human in making timely decisions or in solving complex problem in risky or uncertainty context. A KBS is composed of a knowledge base and an inference mechanism. It operates by storing in its

knowledge base, sentences of semantic information, using the inference mechanism to infer new sentences, and making decisions based on these inferences [Russell and Norvig, 2016].

Moreover, there is a functional need for information and knowledge to be understandable by humans and machines. By doing this it becomes possible to achieve machine reasoning (inference) by applying rules and formal logic to available data in order to offer higher-order deductions. In the course of making decisions, unexpected events emerge and transform the decision space. Sometimes variables in a decision space reach a threshold and the status of the current problem, or decision changes or takes on new characteristics. Such a transformation may require new kinds of information, new expert opinion and judgment, and new information assessment and knowledge-validation processes to make sure a decision will be effective and match the new situation.

In other words, knowledge engineering is the process of understanding and then representing human knowledge in data structures, semantic models (conceptual diagrams of the data as it relates to the real world) and heuristics (AI rules). Expert systems and algorithms are examples that form the basis of the representation and application of this knowledge. The knowledge engineering process includes a) knowledge acquisition, representation, and validation; b) inferencing; and c) explanation and justification.

E.2. Knowledge Engineering Taxonomy

Domain (artificial intelligence)

- [ISO/IEC 2382, 2015] Specific field of knowledge or expertise.

Domain knowledge

- [ISO/IEC 2382, 2015] Knowledge accumulated in a particular domain.

Expert system

- [ISO/IEC 2382, 2015] Computer system that provides for expertly solving problems in a given field or application area by drawing inferences from a knowledge base developed from human expertise. The term "expert system" is sometimes used synonymously with "knowledge-based system", though it is usually taken to emphasize expert knowledge. Some expert systems are able to improve their knowledge base and develop new inference rules based on their experience with previous problems.

Knowledge acquisition

- [ISO/IEC 2382, 2015] Process of locating, collecting, and refining knowledge and converting it into a form that can be further processed by a knowledge-based system. Knowledge acquisition normally implies the intervention of a knowledge engineer, but it is also an important component of machine learning.

Knowledge Base

- [ISO/IEC 2382, 2015] Database that contains inference rules and information about human experience and expertise in a domain. In self-improving systems, the knowledge base additionally contains information resulting from the solution of previously encountered problems.

Knowledge Engineering

- [ISO/IEC 2382, 2015] Discipline concerned with acquiring knowledge from domain experts and other knowledge sources and incorporating it into a knowledge base. The term "knowledge engineering" sometimes refers particularly to the art of designing, building, and maintaining expert systems and other knowledge-based systems.

Knowledge graph

- [ICF, 2018] Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities.

Knowledge representation

- [ISO/IEC 2382, 2015] Process or result of encoding and storing knowledge in a knowledge base

Knowledge-based methods

- [Scharei et al., 2020] Knowledge-based methods are based on ontologies and huge databases of notions, information, and rules.

Knowledge-Based System

- [ISO/IEC 2382, 2015] Information processing system that provides for solving problems in a particular domain or application area by drawing inferences from a knowledge base. The term "knowledge-based system" is sometimes used synonymously with "expert system", which is usually restricted to expert knowledge. Some knowledge-based systems have learning capabilities.

Chapter F

Algorithm Engineering

F.1. Introduction

Why addressing engineering at algorithm level? Let us cite the main reasons. First, if the algorithm study is forgotten or shortened, software may undergo a significant impact. Second, not considering some coarse-grain embedded constraints (e.g. memory size) may lead to fully rethinking for instance a software function that cannot fit the required SWaP (Size Weight and Power) constraints; in that case, the loop back from the software engineer to the algorithm one is very expensive. Third, because Artificial Intelligence (AI) brings a new complexity. Indeed, if most of algorithms issues can be solved at system / software level because they rely on explicit requirements to be verified and validated, this is not always the case for AI algorithms. Last but not least, if a software system has been developed according to some given assumptions, directly modifying the software implementation (without considering the algorithm level again) with respect to a new assumption may lead a disaster since the algorithm choice may be completely different.

Traditionally, algorithms have been studied from the perspective of mathematical algorithm theory, where the analysis of algorithms for combinatorial problems focused on the theoretical analysis of asymptotic worst-case run-times. As a consequence, the development of asymptotically faster algorithms or the improvement of existing ones was a major aim. Sophisticated algorithms and data structures have been developed and new theoretical results were achieved for many problems. With the renewal introduced by AI, the AI algorithm deployment in critical system has caused increasing gaps between theory and practice.

In fact, Algorithm Engineering [Müller-Hannemann and Schirra, 2001] (AE) appeared before the eighties on a theoretical basis (e.g. paper study on complexity). The research field of algorithm engineering copes with these problems and intends to bridge the gap between the efficient algorithms developed in algorithmic theory and the algorithms used by practitioners, aiming to keep the advantages of theoretical treatment: generality, reliability, and predictability from

performance guarantees.

> *"Algorithm Engineering refers to the process required to transform a pencil-and-paper algorithm into a robust, efficient, well tested, and easily usable implementation. Thus it encompasses a number of topics, from modeling cache behavior to the principles of good software engineering; its main focus, however, is experimentation."* - [Sanders, 2009]

Note also that addressed algorithms in this period rather were operational research ones included into knowledge based AI today.

Just like software engineering, AE is not a straight line process. Ideally, one would design an algorithm, implement it, and use it. Thus, the core is a cycle driven by hypotheses. It consist of design, analysis, implementation and experimental evaluation of practicable algorithms. Realistic models for both computers and applications, as well as algorithm libraries and collections of real input data allow a close coupling to applications. The underlying philosophy is depicted on figure F.1.

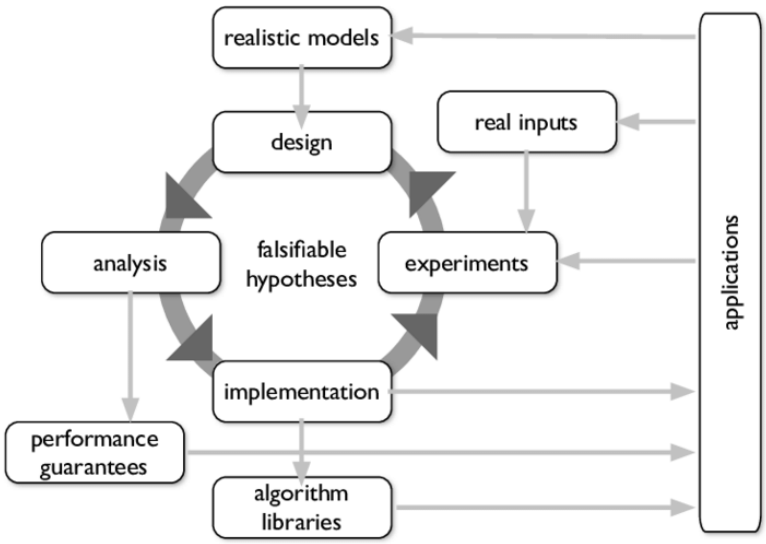

Figure F.1: Basic elements of the algorithm engineering methodology [Sanders, 2009]

The algorithm engineering cycle [Sanders, 2009] proposes multiple iterations over its sub-steps (see figure F.1). It focuses not so much on software engineering aspects but rather on algorithmic development. We usually start with some specific application (1) in mind and try to find a realistic model (2) for it such that the solutions we will obtain match the requirements as well as possible. The main cycle starts with an initial algorithmic design (3) on how to solve this model. Based on this design, we analyze (4) the algorithm from the theoretical point of view, e.g., to obtain performance guarantees (5) such as asymptotic run-time, approximation ratios, etc. These latter two steps to find and analyze an algorithm for a given model are essentially the steps traditionally performed by algorithmic theoreticians in the pen-and-paper. From a link with actual applications, it shows a cycle that might imply a revision of the modeling after some

experimentation. It is an incremental and pragmatic way to build algorithms with respect to reality.

To match industrial needs especially induced by critical systems, the link with practice has to be kept to take into account the sensitivity to real world and its underlying constraints but algorithm engineering must also consider data-driven AI and hybrid AI now. Indeed the current algorithm engineering focuses neither on data-driven AI nor on industrial constraints (e.g., qualification) for instance till now. Moreover, considering a bit more the hardware architecture constraints early will be helpful for software engineering. Thus, it becomes necessary to revisit Algorithm Engineering approach in order to be able to deploy AI in a sound manner within critical systems.

[Confiance.ai, 2021a, Algorithm Engineering State of the Art] gives an overview of algorithm engineering and its application to Artificial intelligence.

F.2. Algorithm Engineering Taxonomy

A posteriori-provability

- [Mamalet et al., 2021] The desired property is verified on the model after training. This approach may also rely on some assumptions on the ML algorithm (e.g. the architecture, the size of the network, the activation function type for a NN...), but these assumptions depends on the problem

A priori-provability or by-design provability

- [Mamalet et al., 2021] The desired property is mathematically ÒtransferableÓ as a design constraint to the ML algorithm. Then, to prove the property, it is necessary to demonstrate the validity of this transfer (i.e., if the design constraint is satisfied then the property holds on the model) and to demonstrate compliance with the design constraint.

Accuracy

- [ISO/IEC/IEEE 24765, 2017]
  1. A qualitative assessment of correctness, or freedom from error.
  2. A quantitative measure of the magnitude of error.
  3. Within the quality management system, accuracy is an assessment of correctness.

Algorithm

- [ISO/IEC/IEEE 24765, 2017]
  1. A finite set of well-defined rules for the solution of a problem in a finite number of steps.
  2. A sequence of operations for performing a specific task.
  3. A finite ordered set of well-defined rules for the solution of a problem.
- [ISO/IEC 11557, 1992] Set of rules for transforming the logical representation of data.

Algorithm Engineering

- [Albers et al., 2009, Demetrescu et al., 2004] The whole process of designing, analyzing, implementing, tuning, debugging and experimentally evaluating computer programs for solving algorithmic problems.
- [Albers et al., 2009, Demetrescu et al., 2004] Methodologies and tools for developing and engineering efficient algorithmic codes and integrating and reinforcing traditional theoretical approaches for the design and analysis of algorithms and data structures.

Correctness

- [ISO/IEC/IEEE 24765, 2017]
  1. The degree to which a system or component is free from faults in its specification, design, and implementation.
  2. The degree to which software, documentation, or other items meet specified requirements.
  3. The degree to which software, documentation, or other items meet user needs and expectations, whether specified or not.
- [ISO/IEC/IEEE 24765, 2017] Degree to which a system or component is free from faults in its specification, design, and implementation.

Dependability

- [Avizienis et al., 2004] The ability to deliver service that can justifiably be trusted. It entails Availability: readiness for correct service; Reliability: continuity of correct service; Safety: absence of catastrophic consequences on the user(s) and the environment; Confidentiality: absence of unauthorized disclosure of information; Integrity: absence of improper system alterations; Maintainability: ability to undergo modifications, and repairs. Security: the concurrent existence of availability for authorized users only, confidentiality, and integrity (with improper meaning unauthorized here)
- [High-Level Expert Group on Artificial Intelligence, 2019] Ability to deliver services that can justifiably be trusted.
- [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required

Effectiveness

- [ISO 9000, 2015] Extent to which planned activities are realized and planned results achieved.

Efficiency

- [ISO 9000, 2015] Relationship between the results achieved and the resources used.
- Relationship between the results achieved and the resources used. Resources expended in relation to the accuracy and completeness with which users achieve goals

Indicator

- [ISO/IEC 27000, 2018] Measure that provides an estimate or evaluation.

Integrity

- [ISO/IEC 27000, 2018] Property of accuracy and completeness.

Chapter G

Software Engineering

G.1. Introduction

Engineering techniques are used to inform the software development process which involves the definition, implementation, assessment, measurement, management, change, and improvement of the software life cycle process itself. It heavily uses software configuration management which is about systematically controlling changes to the configuration, and maintaining the integrity and traceability of the configuration and code throughout the system life cycle. Modern processes use software versioning.

Thus, **Software engineering** is the branch of computer science that deals with the design, development, testing, and maintenance of software applications. Software engineers apply engineering principles and knowledge of programming languages to build software solutions for end users. AI has changed the landscape of software development entirely. With AI-powered tools, developers can now create complex programs and applications with little to no coding required. Since recently, there was no relationship between artificial intelligence and software engineering due to no interaction with the respective communities. Recently, in the last few years, there is a subtle relationship between these two because multiple researchers have started applying tools of artificial intelligence into software engineering and vice-versa.

The objective of Confiance.ai is to revisit the practices of software engineers in light of AI (including machine learning techniques). However, the usual concepts of software engineering remain valid.

G.2. Software Engineering Taxonomy

Acceptance Criteria

- [ISO/IEC/IEEE 24765, 2017]

    1. Criteria that a system or component must satisfy in order to be accepted by a user, customer, or other authorized entity
- A set of conditions that is required to be met before deliverables are accepted

Acceptance test

- [ISO/IEC/IEEE 24765, 2017] Test of a system or functional unit usually performed by the purchaser on his premises after installation with the participation of the vendor to ensure that the contractual requirements are met.

Accountability

- [ISO/IEC TR 24028, 2020] Property that ensures that the actions of an entity may be traced uniquely to that entity
- [ISO 7498-2, 1989] For systems, accountability is a property that ensures that actions of an entity can be traced uniquely to the entity.
- [ISO/IEC 25010, 2011a] Degree to which the actions of an entity can be traced uniquely to the entity.
- [EASA, 2021] Accountability refers to the idea that one is responsible for their action Ð and as a corollary their consequences Ð and must be able to explain their aims, motivations, and reasons.
- [ISO/IEC 25010, 2011a] Degree to which a product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments.

Assessment

- [CENELEC EN 50126, 2011] The undertaking of an investigation in order to arrive at a judgement, based on evidence, of the suitability of a product.
- [ISO/IEC 21827, 2008] Verification of a product, system or service against a standard using the corresponding assessment method to establish compliance and determine the assurance.

Asset

- [ISO/IEC 21827, 2008] Anything that has value to an organization.

Audit

- [CENELEC EN 50126, 2011] A systematic and independent examination to determine whether the procedures specific to the requirements of a product comply with the planned arrangements, are implemented effectively and are suitable to achieve the specified objectives.

- [ISO/IEC 27000, 2018] A systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled.
  - Note 1: An audit can be an internal audit (first party) or an external audit (second party or third party), and it can be a combined audit (combining two or more disciplines).
  - Note 2: An internal audit is conducted by the organization itself, or by an external party on its behalf.

Auditability

- [Mamalet et al., 2021] The extent to which an independent examination of the development and verification process of the system can be performed

Authenticity

- [ISO/IEC 25010, 2011a] Degree to which the identity of a subject or resource can be proved to be the one claimed assets

Availability

- [EN 50129, 2018] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.

Availability

- [ISO/IEC 27000, 2018] Property of being accessible and usable on demand by an authorized entity
- [EN 50129, 2018] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component is operational and accessible when required for use

Black box

- [ISO/IEC/IEEE 24765, 2017]
  1. A system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or implementation are unknown or irrelevant.
  2. Pertaining to an approach that treats a system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or irrelevant.

Capacity

- [ISO/IEC 25010, 2011a] Degree to which the maximum limits of the product or system, parameter meet requirements.

Certification

- [ISO/IEC/IEEE 24765, 2017]
  1. Third-party attestation related to products, processes, systems, or persons.
  2. A written guarantee that a system or component complies with its specified requirements and is acceptable for operational use.
  3. Formal demonstration that a system or component complies with its specified requirements and is acceptable for operational use.
  4. Process of confirming that a system or component complies with its specified requirements and is acceptable for operational use.

Certification Credit

- [RTCA DO-297, 2005] Acceptance by the certification authority that a process, product or demonstration satisfies a certification requirement.

Commercial Off-The-Shelf

- [CENELEC EN 50128, 2020] Software defined by market-driven need, commercially available and whose fitness for purpose has been demonstrated by a broad spectrum of commercial users.

Compliance

- [CENELEC EN 50126, 2011] A demonstration that a characteristic or property of a product satisfies the stated requirements.

Component

- [Szyperski et al., 2002] A basic building-block for systems with well-defined interfaces, behavior and explicit context dependencies only.
  - A component can be deployed independently. That is, it implements a clear function.
  - A component can be composed with other components into systems, sub-system or new components.
  - A component can exist in the form of software or hardware or a combination of both.
- [RTCA DO-297, 2005] A self-contained hardware or software part, database, or combination thereof that may be configuration controlled.

Configuration Management

- [CENELEC EN 50126, 2011] A discipline applying technical and administrative direction and surveillance to identify and document the functional and physical characteristics of a configuration item, control change to those characteristics, record and report change processing and implementation status and verify compliance with specified requirements.
- [RTCA DO-297, 2005] A discipline applying technical and administrative direction and surveillance to (a) identify and record the functional and physical characteristics of a configuration item, (b) control changes to those characteristics, and (c) record and report change control processing and implementation status.

Control

- [IEEE 7000, 2021] Having control of a machine means having
  1. Cognitive control in terms of being informed about what is going on in the computing environment,
  2. Decisional control in terms of having choices over what is going on in one s networked environment,
  3. Behavioral control in terms of receiving feedback on one s actions/choices taken.

  Related and Opposing values
  - Related values: Human responsibility, governance, usability, portability,logic, sense of accomplishment, moderation.
  - Opposing values: Trust, accountability to stakeholders; imagination, reminding, obedience.

Controllability

- [ISO 26262-1, 2018] Ability to avoid a specified harm or damage through the timely reactions of the persons involved, possibly with support from external measures.
- [ISO 26262-1, 2011] Ability to avoid a specified harm or damage through the timely reactions of the persons involved, possibly with support from external measures

Correctness

- [ISO/IEC/IEEE 24765, 2017]
  1. The degree to which a system or component is free from faults in its specification, design, and implementation.
  2. The degree to which software, documentation, or other items meet specified requirements.
  3. The degree to which software, documentation, or other items meet user needs and expectations, whether specified or not
- [Holloway, 2019] The implementation is correct with respect to its defined intended behavior, under foreseeable operating conditions.
- [ISO/IEC/IEEE 24765, 2017] Degree to which a system or component is free from faults in its specification, design, and implementation.

Criticality

- [ISO/IEC/IEEE 24765, 2017] Degree of impact that a requirement, module, error, fault, failure, or other item has on the development or operation of a system.

Criticality Level

- [ANSI/UL 4600, 2020] Level categorizing the risk associated with an unmitigated hazard.

Design (verb)

- [ISO/IEC/IEEE 15288, 2015] The process to define the architecture, system elements, interfaces, and other characteristics of a system or system element.

Effectiveness

- [ISO 9000, 2015] Extent to which planned activities are realized and planned results achieved.
- [ISO 9241-210, 2019] Accuracy and completeness with which users achieve specified goals

Efficiency

- [ISO 9000, 2015] Relationship between the results achieved and the resources used.
- Relationship between the results achieved and the resources used. Resources expended in relation to the accuracy and completeness with which users achieve goals

Error

- [Avižienis et al., 2004] An error is defined as the part of a system s total state that may lead to a failure

Evidence

- [ISO/TS 21089, 2018] Everything that is used to determine or demonstrate the truth of an assertion

Failure

- [ISO 26262-1, 2018] Elimination of the ability of an element, to perform a function as required.

Fault tolerance

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component operates as intended despite the presence of hardware or software faults

Freedom from risk

- A characteristic that is defined as the degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment.

Functional appropriateness

- [ISO/IEC 25010, 2011a] Degree to which the functions facilitate the accomplishment of specified tasks and objectives

Functional completeness

- [ISO/IEC 25010, 2011a] Degree to which the set of functions covers all the specified tasks and user objectives

Functional correctness

- [ISO/IEC 25010, 2011a] Degree to which a product or system provides the correct results with the needed degree of precision

Functional failure

- [ISO 26262-1, 2018] Absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E [Eletrical / Electronic] systems.

Functional suitability

- [ISO/IEC 25010, 2011a] Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

Global Robustness

- [Mamalet et al., 2021] Ability of the system to perform the intended function in the presence of abnormal or unknown inputs

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Grity

- [ISO/IEC TR 29119-11, 2020] System which, given a particular set of inputs and starting state, will always produce the same set of outputs and final state.

Harm

- [ISO GUIDE 51, 2014] Injury or damage to the health of people or damage to property or the environment.

Hazard

- [ISO GUIDE 51, 2014] Potential source of harm.

Incremental Acceptance

- [RTCA DO-297, 2005] A process for obtaining credit toward approval and certification by accepting or finding that an IMA module, application, and/or off-aircraft IMA system complies with specific requirements. Credit granted for individual tasks contributes to the overall certification goal. N.B. very domain specific.

Indicator

- [ISO/IEC 27000, 2018] Measure that provides an estimate or evaluation.

Installability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency in which a product or system can be successfully installed and/or uninstalled in a specified environment.

Integrity

- [ISO/IEC 27000, 2018] Property of accuracy and completeness.

Integrity

- [ISO/IEC 27000, 2018] Property of protecting the accuracy and completeness of assets.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
- [ISO/IEC TR 24028, 202O] An AI system's respect of sound moral and ethical principles or the assurance that information will not be manipulated in a malicious way by the AI system.
- [ISO/IEC/IEEE 24765, 2010]
    1. Value representing project-unique characteristic, such as complexity, criticality, risk, safety level, security level, desired performance, and reliability, that define the importance of the system, software, or hardware to the user.
    2. Degree to which software complies or must comply with a set of stakeholder-selected software and/or software-based system characteristics defined to reflect the importance of the software to its stakeholders.
    3. Symbolic value representing a degree of compliance within an integrity level scheme.

4. Claim of a system, product, or element that includes limitations on a property's values, the claim's scope of applicability, and the allowable uncertainty regarding the claim's achievement.
5. Required degree of confidence that the system-of-interest meets the associated integrity level claim.

Intended behavior

- [ISO/PAS 21448, 2019] Specified behavior of the intended functionality including interaction with items.

Intended functionality

- [ISO/PAS 21448, 2019] behavior specified for a system.

Intent

- [Holloway, 2019] The defined intended behavior is correct and complete with respect to the desired behavior.

Life cycle

- [ISO/IEC/IEEE 15288, 2015] The evolution of a system, product, service, project or other human-made entity from conception through retirement. A life cycle can be described using an abstract functional model that represents the conceptualization of a need for the system, its realization, utilization, evolution and disposal.
- [ISO/IEC DIS 22989, 2021a] Evolution of a system, product, service, project or other human-made entity, from conception through retirement.

Maintainability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements
- The ability to identify and fix a fault within a software component is what the maintainability characteristic addresses. In other software quality models this characteristic is referenced as supportability. Maintainability is impacted by code readability or complexity as well as modularization. Anything that helps with identifying the cause of a fault and then fixing the fault is the concern of maintainability. Also the ability to verify (or test) a system, i.e. testability, is one of the subcharacteristics of maintainability.
- [Mamalet et al., 2021] Ability of extending/improving a given system while maintaining its compliance with the unchanged requirements..

Maturity

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component meets needs for reliability under normal operation.

Measurability

- [ISO/IEC TS 5723, 2022] Ability to assess an attribute of an entity against a metric Note 1: The word "measurable" is the adjective form of measurability.

Measure

- [ISO/IEC 25024, 2015] Variable to which a value is assigned as the result of measurement Note 1: The term measures is used to refer collectively to base measures, derived measures, and indicators.

Measure of Effectiveness

- [of Defense, 2020, Haskins et al., 2006] The operational measures of success that are closely related to the achievement of the mission or operational objective being evaluated, in the intended operational environment under a specified set of conditions; i.e., how well the solution achieves the intended purpose.

Measure of Performance

- [of Defense, 2020, Haskins et al., 2006] The measures that characterize physical or functional attributes relating to the system operation, measured or estimated under specified testing and/or operational environment conditions.

Measurement

- [ISO/IEC 25024, 2015] Set of operations having the object of determining a value of a measure

Measurement function

- [ISO/IEC 25024, 2015] Algorithm or calculation performed to combine two or more quality measure elements

Misuse

- [ISO/PAS 21448, 2019] Usage of the system by a human in a way not intended by the manufacturer of the system.

Modifiability

- [ISO/IEC 25010, 2011a] Degree to which a product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.

Modularity

- [ISO/IEC 25010, 2011a] Degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.

Monitor

- [Leucker and Schallhart, 2009] A monitor is a device that reads a finite trace and yield a certain verdict.

Non-overridable

- [EASA, 2021] Human has no capability to override the AI-based system s operations.

Non-repudiation

- [ISO/IEC 25010, 2011a] Degree to which actions or events can be proven to have taken place, so that the events or actions cannot be repudiated later.

Output Data (of a module)

- [Confiance.ai, 2021b] Output data are data obtained as the output of a given module of a machine learning workflow no matter what the stage in the workflow. Inlier, Outlier, Novelty, or Infeasible Corner Case Data

Performance efficiency

- [ISO/IEC 25010, 2011a] Performance relative to the amount of resources used under stated conditions

Portability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another

Precision

- [ISO/IEC TR 29119-11, 2020] Performance metric used to evaluate a classifier, which measures the proportion of predicted positives that were correct.

- [ISO/DIS 5725-1, 2020] Closeness of agreement between independent test results obtained under stipulated conditions
  - Note 1: Precision depends only on the distribution of random errors and does not relate to the true value or the specified value.
  - Note 2: The measure of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results. Less precision is reflected by a larger standard deviation.
  - Note 3: Quantitative measures of precision depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme conditions.

Process

- [ISO 9000, 2015] Set of interrelated or interacting activities that use inputs to deliver an ntended result.

Provability

- [Mamalet et al., 2021] The extent to which a set of properties on this algorithm can be guaranteed mathematically.

Quality

- [ISO/TS 13972, 2015] Degree to which all the properties and characteristics of a product, process or service satisfy the requirements which ensue from the purpose for which that product, process or service is to be used.

Quality Assurance

- [ISO/IEC/IEEE 15288, 2015] Part of quality management focused on providing confidence that quality requirements will be fulfilled.
- [Daniels et al., 2002, Haskins et al., 2006] Set of activities throughout the entire project life cycle necessary to provide adequate confidence that a product or service conforms to stakeholder requirements or that a process adheres to established methodology.
- Potential synonyms are: Assurance, Product Assurance, Development Assurance, Design Assurance.
- Related notions:
  - Safety Assurance (Quality Assurance in the scope of safety)
  - Quality Control (Inspection contributing to Quality Assurance)

Quality characteristic

- [ISO/IEC/IEEE 15288, 2015] Inherent characteristic of a product, process, or system related to a requirement.

- [ISO/IEC 25023, 2015] Category of quality attributes that bears on software product or system quality

Quality in use

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, satisfaction, and freedom from risk in specific contexts of use.
  - Note 1: The quality in use of a software product or system can be measured and evaluated by the effect of the target system or software products when used by users of the implemented system or during field testing or prototype testing.
  - Note 2: When quality in use is specified, it relates to specified users meeting their needs to achieve specified goals with effectiveness, efficiency, satisfaction, and freedom from risk in specified contexts of use.

Quality Management

- [ISO/IEC/IEEE 15288, 2015] Coordinated activities to direct and control an organization with regard to quality.

Quality measure

- [ISO/IEC 25024, 2015] Measure that is defined as a measurement function of two or more values of quality measure elements

Quality measure element

- [ISO/IEC 25024, 2015] Measure defined in terms of a property and the measurement method for quantifying it, including optionally the transformation by mathematical function

Quality model

- [ISO/IEC 25024, 2015] Defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality

Recoverability

- [ISO/IEC 25010, 2011a] Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.

Reliability

- [ISO/IEC 27000, 2018] Property of consistent intended behavior and results.

Reliability

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Reliability

- [ISO/IEC TS 5723, 2022] Ability of an item to perform as required, without failure, for a given time interval, under given conditions
- [Aerospace, 2010] The probability that an item will perform a required function under specified conditions, without failure, for a specified period
- [ISO/IEC 25012, 2008b] Ability of an item to perform as required, without failure, for a given time interval, under given condition
- [EASA, 2021] The probability that an item will perform a required function under specified conditions, without failure, for a specified period of time

Replaceability

- [ISO/IEC 25010, 2011a] Degree to which a product can replace another specified software product for the same purpose in the same environment.

Requirement

- [ISO/IEC/IEEE 15288, 2015] Statement that translates or expresses a need and its associated constraints and conditions.

Resilience

- [Mamalet et al., 2021] Ability for a system to continue to operate while an error or a fault has occurred
- [High-Level Expert Group on Artificial Intelligence, 2019] Robustness when facing changes.
- [ISO/IEC TS 5723, 2022] Capability of a system to maintain its functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Respect

- [IEEE 7000, 2021] Respect in human-machine interaction implies that a machine is perceived as attentive and responsive. a/ Attentiveness implies that the machine is perceived as replying in a reasonable amount of time and respecting user privacy. b/ Responsiveness implies that the machine is perceived as applying appropriate criteria in its decisions and made explicit to the user and that it is perceived as acting fairly and politely.

- Related and Opposing values
  - Related values: Politeness, courtesy, respect for environment and natural habitat, respect for information and confidentiality, respect for norms, reputation.
  - Opposing values: Self-esteem, maleficence.

Responsibility

- [ISO/IEC 38500, 2015] Obligation to act and take decisions to achieve required outcomes.

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Restricted Operation Domain

- [Colwell et al., 2018] The specific conditions under which a given driving automation system or feature thereof is currently able to function, including, but not limited to, driving modes.

Reusability

- [ISO/IEC 25010, 2011a] Degree to which an asset can be used in more than one system, or in building other assets

Risk

- [CENELEC EN 50126, 2011] The probable rate of occurrence of a hazard causing harm and the degree of severity of the harm.
- [DEFSTAN 00-56(PT1)/7, 2017] Combination of the likelihood of harm and the severity of that harm or long term damage to health.
- [ANSI/UL 4600, 2020] A combination of the probability of occurrence of a loss event and the severity of that loss event.
- [ISO/IEC 27000, 2018] Effect of uncertainty on objectives.
  - Note 1: An effect is a deviation from the expected Ñ positive or negative.
  - Note 2: Uncertainty is the state, even partial, of deficiency of information related to, understanding or knowledge of, an event, its consequence, or likelihood.
  - Note 3: Risk is often characterized by reference to potential events (as defined in ISO-Guide 73:2009, 3.5.1.3) and consequences (as defined in ISO-Guide 73:2009, 3.6.1.3), or a combination of these.
  - Note 4: Risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood (as defined in ISO-Guide 73:2009, 3.6.1.1) of occurrence.

  - Note 5: In the context of information security management systems, information security risks can be expressed as effect of uncertainty on information security objectives.
  - Note 6: Information security risk is associated with the potential that threats will exploit vulnerabilities of an information asset or group of information assets and thereby cause harm to an organization.

Risk Assessment

- [ISO/IEC 27000, 2018] Overall process of risk identification, risk analysis and risk evaluation.

Risk management process

- [ISO GUIDE 73, 2009] Systematic application of management policies, procedures and practices to the activities of communicating, consulting, establishing the context, and identifying, analyzing, evaluating, treating, monitoring and reviewing risk.

Risk mitigation

- [ISO 22300, 2021] Lessening or minimizing of the adverse impacts of a hazardous event.

Robustness

- [ISO/IEC DIS 22989, 2021a, ISO/IEC TR 24029-1, 2021] Ability of a system to maintain its level of performance under a variety of circumstances.
- [Mamalet et al., 2021] (Global) Ability of the system to perform the intended function in the presence of abnormal or unknown inputs / (Local) The extent to which the system provides equivalent responses for similar inputs.
- [EASA, 2021] For an input varying in a region of the state space, the system is producing the same outputs.
- [Gehr et al., 2018] Local robustness (or robustness, for short) requires that all samples in the neighborhood of a given input are classified with the same label.

Runtime monitor (runtime verification)

- [Cassar et al., 2017] Runtime Monitoring is a lightweight and dynamic verification technique that involves observing the internal operations of a software system and/or its interactions with other external entities, with the aim of determining whether the system satisfies or violates a correctness specification.

Satisfaction

- [ISO/IEC 25022, 2016] Degree to which user needs are satisfied when a product or system is used in a specified context of use.

  - Note 1: For a user who does not directly interact with the product or system, only purpose accomplishment and trust are relevant.
  - Note 2: Satisfaction is the user s response to interaction with the product or system, and includes attitudes towards use of the product.
  - Note 3: Users include: primary users who interact with the system to achieve the primary goals, secondary users who provide support, and indirect users who receive output, but do not interact with the system.
  - Note 4: In this International Standard, user's needs include their desires and expectations associated with use of a product, system, or service. Exceeding desires and expectations is a means of significantly increasing satisfaction and improving the user experience.

Software Assurance

- [Hinchey et al., 2006] Software Assurance is the planned and systematic set of activities that ensures that software processes and products conform to requirements, standards and procedures

Software engineering

- [ISO/IEC 2382, 2015] Systematic application of scientific and technological knowledge, methods, and experience to the design, implementation, testing, and documentation of software to optimize its production, support, and quality.

Specifiability

- [Mamalet et al., 2021] The extent to which the system can be correctly and completely described through a list of requirements (such as stakeholder requirements, "black box" requirements, or "white box" requirements).
- Requirement: an identifiable element of a function specification that can be validated and against which an implementation can be verified.

System-Dependent Data Quality

- [ISO/IEC 25012, 2008a] Degree to which data quality is reached and preserved within a computer system when data is used under specified conditions.

Test Dataset

- [Confiance.ai, 2021b] A Dataset that is only composed of Observations that will be used only for evaluating the ML Model performance under operational configuration. Test Dataset must be fully independent of Training Dataset and Validation Dataset. No Observation from this set can be part of these two Datasets.

Testability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.

Time behavior

- [ISO/IEC 25010, 2011a] Degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements

Traceability

- [IEEE 610.12-1990, 2002] (1) The degree to which a relationship can be established between, two or more products of the development process, especially products having a predecessor, successor, or master-subordinate relationship to one another; for example, the degree to which the requirements and design of a given software component match. (2) The degree to which each element in a software development product establishes its reason for existing; for example, the degree to which each element in a bubble chart references the requirement that it satisfies.
- [EASA, 2020] An association between artifacts, such as between process outputs or between an output and its originating process
- [EASA, 2021] The ability to track the journey of a data input through all stages of sampling, labelling, processing and decision-making

Transparency

- [ISO/IEC DIS 22989, 2021b] <organization> Property of an organization that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner. <system> Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability.
- [Arrieta et al., 2020] A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models are divided into three categories: simulatable models, decomposable models and algorithmically transparent models.
- [Brundage et al., 2020] Making information about the characteristics of an AI devel oper s operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.
- [IEEE 7000, 2021] Transparency means that information provided about a system is meaningful, useful, accessible, comprehensive, and truthful. A/ Meaningful means that information about a system should be relevant for users concern or user control. B/ Usefulness of information implies that consumers can act upon it and make choices easily, acting upon

the information provided to them. C/ Accessible means that it is possible to easily obtain and retrieve the relevant information in a machine-readable or other way whether through state-of-the-art electronic channels or via constrained devices or constrained networks. D/ Comprehensive means that information about a system should be asy to read and understand for ordinary people and ot require any expert knowledge. E/ Truthful means that information about a system accurately reflects a system s or system landscape s activities, such as data processing and data sharing practices. The information should be up to date and ritten in plain language that is clear and direct. It should not mislead users in any way, hide information, or give half-truth about practices.

- [ISO/IEC DIS 22989, 2021b] Property of a system that appropriate information about the system is communicated to relevant stakeholders.

Triggering event

- [ISO/PAS 21448, 2019] Specific conditions of a driving scenario that serve as an initiator for a subsequent system reaction possibly leading to a hazardous event.

Usability

- [ISO 9241-210, 2019] Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use
  - The specified users, goals and context of use refer to the particular combination of users, goals and context of use for which usability is being considered.
  - The word usability is also used as a qualifier to refer to the design knowledge, competencies, activities and design attributes that contribute to usability, such as usability expertise, usability professional, usability engineering, usability method, usability evaluation, usability heuristic.
- [ISO/IEC 25010, 2011a] Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

User

- [ISO/IEC/IEEE 15288, 2015] Individual or group that interacts with a system or benefits from a system during its utilization.

User controllability

- Involved individual's possibility of avoiding harm in the situation that is putting him/her at risk

User error protection

- [ISO/IEC 25010, 2011a] Degree to which a product or system protects users against making errors

User interface aesthetics

- [ISO/IEC 25010, 2011a] Degree to which a user interface enables pleasing and satisfying interaction for the user.

Validation

- [Aerospace, 2010] The determination that the requirements for a product are correct and complete. (Are we building the right aircraft / system / function / item?
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled (the right system was built).

Validation Dataset

- [Confiance.ai, 2021b] A Dataset that is only composed Observations that will be used only for evaluating the generalization capabilities of the ML Model and choosing / tuning Hyperparameters values. In that context, Validation Dataset is composed of Observations that are not already in the Training Dataset.

Value

- [ISO 10303-1, 2021] Belief(s) an organization adheres to and the standards that it seeks to observe.

Verifiability

- [Mamalet et al., 2021] Ability to evaluate an implementation of requirements to determine that they have been met (adapted from ARP4754A)

Verification

- [Aerospace, 2010] The evaluation of an implementation of requirements to determine that they have been met. (Did we build the aircraft / system/ function / item right?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled (the system was built right).

Verifliable

- [ISO/IEC/IEEE 15288, 2015] Can be checked for correctness by a person or tool

Vulnerability

- [ISO/IEC 27000, 2018] Weakness of an asset or control that can be exploited by one or more threats.
- [ISO GUIDE 73, 2009] Intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence.

Chapter H

Safety Engineering

H.1. Introduction

H.2. Safety Engineering Taxonomy

A posteriori-provability

- [Mamalet et al., 2021] The desired property is verified on the model after training. This approach may also rely on some assumptions on the ML algorithm (e.g. the architecture, the size of the network, the activation function type for a NN...), but these assumptions depends on the problem

A priori-provability or by-design provability

- [Mamalet et al., 2021] The desired property is mathematically transferable as a design constraint to the ML algorithm. Then, to prove the property, it is necessary to demonstrate the validity of this transfer (i.e., if the design constraint is satisfied then the property holds on the model) and to demonstrate compliance with the design constraint.

Acceptance Criteria

- [ISO/IEC/IEEE 24765, 2017]
  1. Criteria that a system or component must satisfy in order to be accepted by a user, customer, or other authorized entity
  2. A set of conditions that is required to be met before deliverables are accepted

Acceptance test

- [ISO/IEC/IEEE 24765, 2017] Test of a system or functional unit usually performed by the purchaser on his premises after installation with the participation of the vendor to ensure

that the contractual requirements are met.

Adversarial Attack

- [Barreno et al., 2006] Action targeting a learning system to cause malfunction.

AI safety risk

- [EASA, 2021] The AI safety risk mitigation building block considers that we may not always be able to open the AI black box to the extent required and that the safety risk may need to be addressed to deal with the inherent uncertainty of AI.

Assurance

- [ISO/TS 21089, 2018] Development, documentation, testing, procedural and operational activities carried out to ensure a system's services do in fact provide the claimed level of function, performance and usability

Assurance Case

- [Rushby, 2015] An assurance case provides an argument to justify certain claims about a system, based on evidence concerning both the system and the environment in which it operates. The claims can be about any system property, such as reliability or security, and thereby generalize the previously established notion of a safety case, where the claim is always about safety.
- [ISO/IEC/IEEE 15026-1, 2019] Reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)
- An assurance case is a structured argument, supported by evidence, intended to justify that a system is acceptably assured relative to a concern (such as safety or security) in the intended operating environment.

Audit

- [CENELEC EN 50126, 2011] A systematic and independent examination to determine whether the procedures specific to the requirements of a product comply with the planned arrangements, are implemented effectively and are suitable to achieve the specified objectives.
- [ISO/IEC 27000, 2018] A systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled. Note 1: An audit can be an internal audit (first party) or an external audit (second party or third party), and it can be a combined audit (combining two or more disciplines). Note 2: An internal audit is conducted by the organization itself, or by an external party on its behalf.

Auditability

- [Mamalet et al., 2021] The extent to which an independent examination of the development and verification process of the system can be performed

Confidentiality

- [ISO/IEC 25010, 2011a] Degree to which the prototype ensures that data are accessible only to those authorized to have access.

Dependability

- [Avizienis et al., 2004] The ability to deliver service that can justifiably be trusted. It entails:
  - Availability: readiness for correct service;
  - Reliability: continuity of correct service;
  - Safety: absence of catastrophic consequences on the user(s) and the environment;
  - Confidentiality: absence of unauthorized disclosure of information;
  - Integrity: absence of improper system alterations;
  - Maintainability: ability to undergo modifications, and repairs.
  - Security: the concurrent existence of availability for authorized users only, confidentiality, and integrity (with improper meaning unauthorized here)
- [High-Level Expert Group on Artificial Intelligence, 2019] Ability to deliver services that can justifiably be trusted.
- [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required
  - Note 1: Dependability includes availability, reliability, recoverability, maintainability, and maintenance support performance, and, in some cases, other characteristics such as durability, safety and security.
  - Note 2: Dependability is used as a collective term for the time-related quality characteristics of an item.
  - [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required

Dynamic Assurance Case

- [ASAADI et al., 2020] Generic framework to provide justified confidence in the capabilities of autonomous systems embedding machine learning-based components.

Error

- [Avižienis et al., 2004] An error is defined as the part of a system s total state that may lead to a failure

Evidence

- [ISO/TS 21089, 2018] Everything that is used to determine or demonstrate the truth of an assertion

Failure

- [ISO 26262-1, 2018] Elimination of the ability of an element, to perform a function as required.

Fault tolerance

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component operates as intended despite the presence of hardware or software faults

Freedom from risk

- A characteristic that is defined as the degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment.

Functional failure

- [ISO 26262-1, 2018] Absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E [Eletrical / Electronic] systems.

Global Robustness

- [Mamalet et al., 2021] Ability of the system to perform the intended function in the presence of abnormal or unknown inputs

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Harm

- [ISO GUIDE 51, 2014] Injury or damage to the health of people or damage to property or the environment.

Hazard

- [ISO GUIDE 51, 2014] Potential source of harm.

High-risk AI System

- [European Commission, 2021] AI systems that create a high risk to the health and safety or fundamental rights of natural persons.

Incremental Acceptance

- [RTCA DO-297, 2005] A process for obtaining credit toward approval and certification by accepting or finding that an IMA module, application, and/or off-aircraft IMA system complies with specific requirements. Credit granted for individual tasks contributes to the overall certification goal. N.B. very domain specific.

Indicator

- [ISO/IEC 27000, 2018] Measure that provides an estimate or evaluation.

Innocuity

- [Holloway, 2019] Any part of the implementation that is not required by the defined intended behavior has no unacceptable impact.

Installability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency in which a product or system can be successfully installed and/or uninstalled in a specified environment.

Integrity

- [ISO/IEC 27000, 2018] Property of protecting the accuracy and completeness of assets.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
- [ISO/IEC TR 24028, 202O] An AI system s respect of sound moral and ethical principles or the assurance that information will not be manipulated in a malicious way by the AI system.

Integrity level

- [ISO/IEC/IEEE 24765, 2017]
  - Value representing project-unique characteristic, such as complexity, criticality, risk, safety level, security level, desired performance, and reliability, that define the importance of the system, software, or hardware to the user.
  - Degree to which software complies or must comply with a set of stakeholder-selected software and/or software-based system characteristics defined to reflect the importance of the software to its stakeholders.
  - Symbolic value representing a degree of compliance within an integrity level scheme.

  - Claim of a system, product, or element that includes limitations on a property's values, the claim's scope of applicability, and the allowable uncertainty regarding the claim's achievement.
  - Required degree of confidence that the system-of-interest meets the associated integrity level claim.

Intended behavior

- [ISO/PAS 21448, 2019] Specified behavior of the intended functionality including interaction with items.

Intended functionality

- [ISO/PAS 21448, 2019] behavior specified for a system.

Intent

- [Holloway, 2019] The defined intended behavior is correct and complete with respect to the desired behavior.

Level of risk

- [ISO/IEC 27000, 2018] Magnitude of a risk expressed in terms of the combination of consequences and their likelihood.

Life cycle

- [ISO/IEC/IEEE 15288, 2015] The evolution of a system, product, service, project or other human-made entity from conception through retirement. A life cycle can be described using an abstract functional model that represents the conceptualization of a need for the system, its realization, utilization, evolution and disposal.
- [ISO/IEC DIS 22989, 2021a] Evolution of a system, product, service, project or other human-made entity, from conception through retirement.

Local Robustness

- [Mamalet et al., 2021] The extent to which the system provides equivalent responses for similar inputs.

Maintainability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements

- The ability to identify and fix a fault within a software component is what the maintainability characteristic addresses. In other software quality models this characteristic is referenced as supportability. Maintainability is impacted by code readability or complexity as well as modularization. Anything that helps with identifying the cause of a fault and then fixing the fault is the concern of maintainability. Also the ability to verify (or test) a system, i.e. testability, is one of the subcharacteristics of maintainability.
- [Mamalet et al., 2021] Ability of extending/improving a given system while maintaining its compliance with the unchanged requirements..

Maturity

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component meets needs for reliability under normal operation.

Mission-critical system

- A system that is essential to the survival of a business or organization.

Misuse

- [ISO/PAS 21448, 2019] Usage of the system by a human in a way not intended by the manufacturer of the system.

Monitor

- [Leucker and Schallhart, 2009] A monitor is a device that reads a finite trace and yield a certain verdict.

Non-overridable

- [EASA, 2021] Human has no capability to override the AI-based system s operations.

Performance efficiency

- [ISO/IEC 25010, 2011a] Performance relative to the amount of resources used under stated conditions

Provability

- [Mamalet et al., 2021] The extent to which a set of properties on this algorithm can be guaranteed mathematically.

Quality

- [ISO/TS 13972, 2015] Degree to which all the properties and characteristics of a product, process or service satisfy the requirements which ensue from the purpose for which that product, process or service is to be used.

Quality Assurance

- [ISO/IEC/IEEE 15288, 2015] Part of quality management focused on providing confidence that quality requirements will be fulfilled.
- [Daniels et al., 2002, Haskins et al., 2006] Set of activities throughout the entire project life cycle necessary to provide adequate confidence that a product or service conforms to stakeholder requirements or that a process adheres to established methodology.
- Potential synonyms are: Assurance, Product Assurance, Development Assurance, Design Assurance.
- Related notions:
  - Safety Assurance (Quality Assurance in the scope of safety)
  - Quality Control (Inspection contributing to Quality Assurance)

Quality characteristic

- [ISO/IEC/IEEE 15288, 2015] Inherent characteristic of a product, process, or system related to a requirement.
- [ISO/IEC 25023, 2015] Category of quality attributes that bears on software product or system quality

Quality in use

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, satisfaction, and freedom from risk in specific contexts of use
  - Note 1: The quality in use of a software product or system can be measured and evaluated by the effect of the target system or software products when used by users of the implemented system or during field testing or prototype testing.
  - Note 2: When quality in use is specified, it relates to specified users meeting their needs to achieve specified goals with effectiveness, efficiency, satisfaction, and freedom from risk in specified contexts of use.

Quality Management

- [ISO/IEC/IEEE 15288, 2015] Coordinated activities to direct and control an organization with regard to quality.

Quality measure

- [ISO/IEC 25024, 2015] Measure that is defined as a measurement function of two or more values of quality measure elements

Quality measure element

- [ISO/IEC 25024, 2015] Measure defined in terms of a property and the measurement method for quantifying it, including optionally the transformation by mathematical function

Quality model

- [ISO/IEC 25024, 2015] Defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality

Recoverability

- [ISO/IEC 25010, 2011a] Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.

Reliability

- [ISO/IEC 27000, 2018] Property of consistent intended behavior and results.

Reliability

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.
- [ISO/IEC TS 5723, 2022] Ability of an item to perform as required, without failure, for a given time interval, under given conditions
- [Aerospace, 2010] The probability that an item will perform a required function under specified conditions, without failure, for a specified period
- Ability of an item to perform as required, without failure, for a given time interval, under given condition
- [EASA, 2021] The probability that an item will perform a required function under specified conditions, without failure, for a specified period of time

Replaceability

- [ISO/IEC 25010, 2011a] Degree to which a product can replace another specified software product for the same purpose in the same environment.

Requirement

- [ISO/IEC/IEEE 15288, 2015] Statement that translates or expresses a need and its associated constraints and conditions.

Resilience

- [Mamalet et al., 2021] Ability for a system to continue to operate while an error or a fault has occurred
- [High-Level Expert Group on Artificial Intelligence, 2019] Robustness when facing changes.
- [ISO/IEC TS 5723, 2022] Capability of a system to maintain its functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Responsibility

- [ISO/IEC 38500, 2015] Obligation to act and take decisions to achieve required outcomes.

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Reusability

- [ISO/IEC 25010, 2011a] Degree to which an asset can be used in more than one system, or in building other assets

Risk

- [CENELEC EN 50126, 2011] The probable rate of occurrence of a hazard causing harm and the degree of severity of the harm.
- [DEFSTAN 00-56(PT1)/7, 2017] Combination of the likelihood of harm and the severity of that harm or long term damage to health.
- [ANSI/UL 4600, 2020] A combination of the probability of occurrence of a loss event and the severity of that loss event.
- [ISO/IEC 27000, 2018] Effect of uncertainty on objectives.
  - Note 1: An effect is a deviation from the expected Ñ positive or negative.
  - Note 2: Uncertainty is the state, even partial, of deficiency of information related to, understanding or knowledge of, an event, its consequence, or likelihood.

  - Note 3: Risk is often characterized by reference to potential events (as defined in ISO-Guide 73:2009, 3.5.1.3) and consequences (as defined in ISO-Guide 73:2009, 3.6.1.3), or a combination of these.
  - Note 4: Risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood (as defined in ISO-Guide 73:2009, 3.6.1.1) of occurrence.
  - Note 5: In the context of information security management systems, information security risks can be expressed as effect of uncertainty on information security objectives.
  - Note 6: Information security risk is associated with the potential that threats will exploit vulnerabilities of an information asset or group of information assets and thereby cause harm to an organization.

Risk Assessment

- [ISO/IEC 27000, 2018] Overall process of risk identification, risk analysis and risk evaluation.

Risk management process

- [ISO GUIDE 73, 2009] Systematic application of management policies, procedures and practices to the activities of communicating, consulting, establishing the context, and identifying, analyzing, evaluating, treating, monitoring and reviewing risk.

Risk mitigation

- [ISO 22300, 2021] Lessening or minimizing of the adverse impacts of a hazardous event.

Robustness

- [ISO/IEC DIS 22989, 2021a, ISO/IEC TR 24029-1, 2021] Ability of a system to maintain its level of performance under a variety of circumstances.
- [Mamalet et al., 2021] (Global) Ability of the system to perform the intended function in the presence of abnormal or unknown inputs / (Local) The extent to which the system provides equivalent responses for similar inputs.
- [EASA, 2021] For an input varying in a region of the state space, the system is producing the same outputs.
- [Gehr et al., 2018] Local robustness (or robustness, for short) requires that all samples in the neighbourhood of a given input are classified with the same label.

Runtime monitor (runtime verification)

- [Cassar et al., 2017] Runtime Monitoring is a lightweight and dynamic verification technique that involves observing the internal operations of a software system and/or its in-

teractions with other external entities, with the aim of determining whether the system satisfies or violates a correctness specification.

Safety

- [ISO/IEC 12207, 2017] Expectation that a system does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.
- [ISO/IEC 25010, 2011b] The ability to have acceptable risk levels in relation with damage to people, companies, software, property, or environment.
- [EN 50129, 2018] Freedom from unacceptable risk of harm.
- [ISO/IEC 12207, 2017] Property of a system such that it does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered
- [ISO GUIDE 51, 2014] Freedom from risk which is not tolerable.

Safety net

- [Green et al., 2011] A safety net is a set of mitigation and protections means aimed at ensuring continuous safe operation in the presence of faults.

Safety Of The Intended Functionality (SOTIF)

- [ISO/PAS 21448, 2019] Absence of unreasonable risk due to hazards resulting from functional insufficiencies of the intended functionality or from reasonably foreseeable misuse by persons.

Safety-critical system

- A system whose failure may result in injury, loss of life or serious environmental damage.

Specifiability

- [Mamalet et al., 2021] The extent to which the system can be correctly and completely described through a list of requirements (such as stakeholder requirements, "black box" requirements, or "white box" requirements). Requirement: an identifiable element of a function specification that can be validated and against which an implementation can be verified.

Testability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.

Time behavior

- [ISO/IEC 25010, 2011a] Degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements

Triggering event

- [ISO/PAS 21448, 2019] Specific conditions of a driving scenario that serve as an initiator for a subsequent system reaction possibly leading to a hazardous event.

Validation

- [Aerospace, 2010] The determination that the requirements for a product are correct and complete. (Are we building the right aircraft / system / function / item?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled (the right system was built).

Verifiability

- [Mamalet et al., 2021] Ability to evaluate an implementation of requirements to determine that they have been met (adapted from ARP4754A)

Verification

- [Aerospace, 2010] The evaluation of an implementation of requirements to determine that they have been met. (Did we build the aircraft / system/ function / item right?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled (the system was built right).

Verifiable

- [ISO/IEC/IEEE 15288, 2015] Can be checked for correctness by a person or tool

Vulnerability

- [ISO/IEC 27000, 2018] Weakness of an asset or control that can be exploited by one or more threats.
- [ISO GUIDE 73, 2009] Intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence.

Chapter I

Cyber-Security Engineering

I.1. Introduction

I.2. Cyber-security Engineering taxonomy

Adversarial - Black-box attack (Zero knowledge attack)

- [Akhtar and Mian, 2018, Bak and Duggirala, 2017, Chakraborty et al., 2018] Attack that assumes no knowledge about the model under attack. The adversary may use context or historical information to infer model vulnerability. The attacker may probe the system to inform system vulnerabilities.

Adversarial - Evasion attack

- [Biggio and Roli, 2018] The attacker manipulates input samples to evade (cause a misclassification) a trained classifier at test time.

Adversarial - Fast Gradient Sign Method (FGSM)

- [Akhtar and Mian, 2018, Bak and Duggirala, 2017] An efficient method for computing an adversarial image perturbation, using the gradient of the cost function. The image is perturbed to increase the loss of the classifier on the resulting image.

Adversarial - Gray-box attack (Limited knowledge attack)

- [Akhtar and Mian, 2018, Bak and Duggirala, 2017, Biggio and Roli, 2018] Attack which assumes partial knowledge about the model under attack (e.g., type of features, or type of training data).

Adversarial - Jacobian-based Saliency Map Attack (JSMA)

- [Akhtar and Mian, 2018, Papernot et al., 2018] An attack that makes optimal miniscule changes to input data until the classifier is fooled or a maximum number of changes is met.

Adversarial - Non-targeted attack (Untargeted attack)

- [Akhtar and Mian, 2018, Chakraborty et al., 2018, Papernot et al., 2018] An attack that causes any misclassification as opposed to causing classification into a specific (incorrect) class. The predicted label of the adversarial example is irrelevant, as long as it is not the correct label.

Adversarial - One Pixel Attack

- [Akhtar and Mian, 2018] An (evasion) attack that alters a single pixel in an image to cause a misclassification.

Adversarial - Poisoning attack

- [Chakraborty et al., 2018, Liu et al., 2018] Aims to increase the number of misclassified samples at test time by injecting a small fraction of carefully designed adversarial samples into the training data. Indirect poisoning manipulates data before any preprocessing, while direct poisoning alters either the data by Data Injection or Data Manipulation, or the model itself by Logic Corruption. Also known as a contamination of the training data. Alternately, also includes tampering with the ML algorithm itself, to compromise the whole learning process.

Adversarial - Real-world attacks

- [Akhtar and Mian, 2018] Attacks successfully executed on existing systems.

Adversarial - Targeted misclassification attack

- [Chakraborty et al., 2018] The adversary tries to produce inputs that force the output of the classification model to be a specific target class. For example, any input image to the classification model will be predicted as a class of images having a Speed Limit sign.

Adversarial - Threat model

- [Biggio and Roli, 2018, Chakraborty et al., 2018],[Papernot et al., 2018] Adversarial goals, knowledge, and capabilities that a system is designed to defend against.

Adversarial - Universal (Adversarial) perturbation

- [Akhtar and Mian, 2018] Perturbation able to fool a given model on any image with high probability.

Adversarial - White-box attack (Perfect knowledge attack)

- [Akhtar and Mian, 2018, Chakraborty et al., 2018, Papernot et al., 2018] Attack that exploits model internal information. It assumes complete knowledge of the targeted model, including its parameter values, architecture, training method, and in some cases its training data as well.

Adversarial Attack

- [Barreno et al., 2006] Action targeting a learning system to cause malfunction.

Adversarial example

- [Akhtar and Mian, 2018, Bak and Duggirala, 2017] ML input sample formed by applying a small but intentionally worst-case perturbation to a clean example, such that the perturbed input causes a learned model to output an incorrect answer.

Adversarial perturbation

- [Akhtar and Mian, 2018, Bak and Duggirala, 2017] The noise added to an input sample to make it an adversarial example.

Adversarial training

- [Chakraborty et al., 2018] Defensive method to increase model robustness by injecting adversarial examples into the training set.

Adversary

- [Akhtar and Mian, 2018, Bak and Duggirala, 2017] The agent who conducts or intends to conduct detrimental activities, perhaps by creating an adversarial example.

Assurance

- [ISO/TS 21089, 2018] Development, documentation, testing, procedural and operational activities carried out to ensure a system's services do in fact provide the claimed level of function, performance and usability

Assurance Case

- [Rushby, 2015] An assurance case provides an argument to justify certain claims about a system, based on evidence concerning both the system and the environment in which it operates. The claims can be about any system property, such as reliability or security, and thereby generalize the previously established notion of a safety case, where the claim is always about safety.

Assurance Case

- [ISO/IEC/IEEE 15026-1, 2019] Reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)

Assurance Case

- [Mansourov and Campara, 2010] An assurance case is a structured argument, supported by evidence, intended to justify that a system is acceptably assured relative to a concern (such as safety or security) in the intended operating environment.

Dependability

- [Avizienis et al., 2004] The ability to deliver service that can justifiably be trusted. It entails:
  - Availability: readiness for correct service;
  - Reliability: continuity of correct service;
  - Safety: absence of catastrophic consequences on the user(s) and the environment;
  - Confidentiality: absence of unauthorized disclosure of information;
  - Integrity: absence of improper system alterations;
  - Maintainability: ability to undergo modifications, and repairs.
  - Security: the concurrent existence of availability for authorized users only, confidentiality, and integrity (with improper meaning unauthorized here)
- [High-Level Expert Group on Artificial Intelligence, 2019] Ability to deliver services that can justifiably be trusted.
- [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required
  - Note 1: Dependability includes availability, reliability, recoverability, maintainability, and maintenance support performance, and, in some cases, other characteristics such as durability, safety and security.
  - Note 2: Dependability is used as a collective term for the time-related quality characteristics of an item.
  - [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required

Generative adversarial networks (GANs)

- [Techopedia, 2018b] A type of construct in neural network technology that is composed of two neural networks: a generative network, that generates samples, and a discriminative network, that tries to detect whether a sample is real or the result of the generator.

Integrity

- [ISO/IEC 27000, 2018] Property of accuracy and completeness.
- [ISO/IEC 27000, 2018] Property of protecting the accuracy and completeness of assets.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
- [ISO/IEC TR 24028, 202O] An AI systems respect of sound moral and ethical principles or the assurance that information will not be manipulated in a malicious way by the AI system.

Integrity level

- [ISO/IEC/IEEE 24765, 2017]
  1. Value representing project-unique characteristic, such as complexity, criticality, risk, safety level, security level, desired performance, and reliability, that define the importance of the system, software, or hardware to the user.
  2. Degree to which software complies or must comply with a set of stakeholder-selected software and/or software-based system characteristics defined to reflect the importance of the software to its stakeholders.
  3. Symbolic value representing a degree of compliance within an integrity level scheme.
  4. Claim of a system, product, or element that includes limitations on a property's values, the claim's scope of applicability, and the allowable uncertainty regarding the claim's achievement.
  5. Required degree of confidence that the system-of-interest meets the associated integrity level claim.

Monitor

- [Leucker and Schallhart, 2009] A monitor is a device that reads a finite trace and yield a certain verdict.

Privacy

- [ISO/IEC 2382, 2015] Freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual.
- [IEEE 7000, 2021] Privacy means that the collection, processing, and dissemination of personal information is done in such a way as to maintain the information self-determination of a data subject.

- Related and Opposing values
  - Related values: Respect for confidentiality, intimacy, anonymity.
  - Opposing values: Transparency,inclusiveness, alerting.

Resilience

- [Mamalet et al., 2021] Ability for a system to continue to operate while an error or a fault has occurred
- [High-Level Expert Group on Artificial Intelligence, 2019] Robustness when facing changes.
- [ISO/IEC TS 5723, 2022] Capability of a system to maintain its functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary

Robustness

- [Mamalet et al., 2021] (Global) Ability of the system to perform the intended function in the presence of abnormal or unknown inputs / (Local) The extent to which the system provides equivalent responses for similar inputs.
- [EASA, 2021] For an input varying in a region of the state space, the system is producing the same outputs.
- [Gehr et al., 2018] Local robustness (or robustness, for short) requires that all samples in the neighborhood of a given input are classified with the same label.

Security

- [ISO/IEC 25010, 2011b] Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.
- [ISO/IEC 23643, 2020] Resistance to intentional, unauthorized act(s) designed to cause harm or damage to a system

Vulnerability

- [ISO/IEC 27000, 2018] Weakness of an asset or control that can be exploited by one or more threats.
- [ISO GUIDE 73, 2009] Intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence.

Chapter J

Cognitive Engineering

J.1. Introduction

J.2. Cognitive Taxonomy

Acceptance Criteria

- [ISO/IEC/IEEE 24765, 2017]
  1. Criteria that a system or component must satisfy in order to be accepted by a user, customer, or other authorized entity
  2. A set of conditions that is required to be met before deliverables are accepted

Acceptance test

- [ISO/IEC/IEEE 24765, 2017] Test of a system or functional unit usually performed by the purchaser on his premises after installation with the participation of the vendor to ensure that the contractual requirements are met.

Availability

- [EN 50129, 2018] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.
- [ISO/IEC 27000, 2018] Property of being accessible and usable on demand by an authorized entity
- [EN 50129, 2018] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component is operational and accessible when required for use.

Completeness of explainability

- [Mamalet et al., 2021] Relates to the capability to describe a phenomenon in such a way that this description can be used to reach a given goal.

Comprehensibility

- [Arrieta et al., 2020] When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. This notion of model comprehensibility stems from the postulates of Michalski, which stated that the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single chunks of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity.
- [Arrieta et al., 2020] when conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.

Context completeness

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used with the required levels of effectiveness, efficiency, satisfaction, and freedom from risk in each of the specified contexts of use.
  - Note 1: Context completeness is a subcharacteristic of context coverage.

Context coverage

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used with effectiveness, efficiency, satisfaction, and freedom from risk in both specified contexts of use and in contexts beyond those initially explicitly identified
  - Note 1: Context of use is relevant to both quality in use and some product quality (sub)characteristics (where it is referred to as specified conditions).

Context of use

- [ISO/IEC 25022, 2016] Users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a system, product or service is used

Control

- [IEEE 7000, 2021] Having control of a machine means having

1. Cognitive control in terms of being informed about what is going on in the computing environment,
2. Decisional control in terms of having choices over what is going on in one s networked environment,
3. Behavioral control in terms of receiving feedback on one s actions/choices taken.

- Related and Opposing values
  - Related values: Human responsibility, governance, usability, portability,logic, sense of accomplishment, moderation.
  - Opposing values: Trust, accountability to stakeholders; imagination, reminding, obedience.

Controllability

- [ISO 26262-1, 2018] Ability to avoid a specified harm or damage through the timely reactions of the persons involved, possibly with support from external measures.

Ethical

- [Cambridge Dictionary, 2020] Morally right (or, alternatively: relating to beliefs about what is morally right and wrong).

Explainability

- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.
- [Phillips et al., 2020] Explanation: Systems deliver accompanying evidence or reason(s) for all outputs.
  - Meaningful: Systems provide explanations that are understandable to individual users.
  - Explanation Accuracy: The explanation correctly reflects the system s process for generating the output.
  - Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.
- [Arrieta et al., 2020] Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.
- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.

Explainable Artificial Intelligence (XAI)

- [Arrieta et al., 2020] An explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Explainable model

- An explanation of a model result is a description of how a model s outcome came to be.

Explanation facility

- [ISO/IEC 2382, 2015] Component of a knowledge-based system that explains how solutions were derived and justifies the steps used in reaching them.

Fairness

- [Stevenson, 2015] Impartial and just treatment or behavior without favoritism or discrimination.
- [High-Level Expert Group on Artificial Intelligence, 2019] Fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as substantive fairness. But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated
- [IEEE 7000, 2021] Fairness has the attributes of systematic discrimination with an absence of bias in reaching reasonable judgments and allowing opportunities.
  Related and Opposing values
  - Related Values: Responsible position on conflicts of interest, tolerance, justice, balance, equality (legal, gender, minority)
  - Opposing values: Bias, suspicion, discrimination, arbitrariness

Functional appropriateness

- [ISO/IEC 25010, 2011a] Degree to which the functions facilitate the accomplishment of specified tasks and objectives

Functional completeness

- [ISO/IEC 25010, 2011a] Degree to which the set of functions covers all the specified tasks and user objectives

Functional suitability

- [ISO/IEC 25010, 2011a] Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

Fundamental rights

- [High-Level Expert Group on Artificial Intelligence, 2019] Human rights enshrined in the EU Treaties, the Charter of Fundamental Rights (the Charter), and international human rights Law.

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Human Cognitive Bias

- [ISO/IEC TR 24027, 2021] Bias that occurs when humans are processing and interpreting information
  - Note 1: human cognitive bias influences judgment and decision-making.

Human Factors

- [ISO/IEC TR 24028, 2020] External environmental, organizational, and job factors which influence behavior, physical or cognitive characteristics, or social behavior, of a person.
- Human factors can have a significant influence on the interaction within, and the functioning of, management systems.

Human-centred design

- [ISO 9241-210, 2019] Approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors/ergonomics and usability knowledge and techniques
  - Note 1: The term human-centered design is used rather than user-centered design in order to emphasize that this document also addresses impacts on a number of stakeholders, not just those typically considered as users. However, in practice, these terms are often used synonymously.
  - Note 2: Usable systems can provide a number of benefits, including improved productivity, enhanced user well-being, avoidance of stress, increased accessibility and reduced risk of harm.

Inclusiveness

- [IEEE 7000, 2021] Inclusiveness in a system means that it is accessible to differently abled users, unbiased in its decisions, and fair to the broadest range of characteristics (especially human characteristics) it may encounter.
  Related and Opposing values
  - Related values: Participation, partnership, solidarity, interdependence,compatibility, accessibility, diversity
  - Opposing values: Control, bias, detachment

Interpretability

- [Arrieta et al., 2020] The ability to explain or to provide the meaning in understandable terms to a human.
- [Mamalet et al., 2021] Relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a human being. It is a basic requirement for an explanation.

Interpretable model

- An interpretable model should provide users with a description of what a stimulus, such as a datapoint or model output, means in context.

Knowledge acquisition

- [ISO/IEC 2382, 2015] Process of locating, collecting, and refining knowledge and converting it into a form that can be further processed by a knowledge-based system.
- Knowledge acquisition normally implies the intervention of a knowledge engineer, but it is also an important component of machine learning.

Perception

- [Scharei et al., 2020] Perception describes knowledge about the environment by transforming raw sensorial inputs into technically processable information. Sensorial inputs can be images, videos, or sound data.

Respect

- [IEEE 7000, 2021] Respect in human-machine interaction implies that a machine is perceived as attentive and responsive.
  1. Attentiveness implies that the machine is perceived as replying in a reasonable amount of time and respecting user privacy.
  2. Responsiveness implies that the machine is perceived as applying appropriate criteria in its decisions and made explicit to the user and that it is perceived as acting fairly and politely

  Related and Opposing values
  - Related values: Politeness, courtesy, respect for environment and natural habitat, respect for information and confidentiality, respect for norms, reputation.
  - Opposing values: Self-esteem, maleficence.

Responsibility

- [ISO/IEC 38500, 2015] Obligation to act and take decisions to achieve required outcomes.

Satisfaction

- [ISO/IEC 25022, 2016] Degree to which user needs are satisfied when a product or system is used in a specified context of use
  - Note 1: For a user who does not directly interact with the product or system, only purpose accomplishment and trust are relevant.
  - Note 2: Satisfaction is the user s response to interaction with the product or system, and includes attitudes towards use of the product.
  - Note 3: Users include: primary users who interact with the system to achieve the primary goals, secondary users who provide support, and indirect users who receive output, but do not interact with the system.
  - Note 4: In this International Standard, user's needs include their desires and expectations associated with use of a product, system, or service. Exceeding desires and expectations is a means of significantly increasing satisfaction and improving the user experience.

Stakeholder

- [ISO/IEC 38500, 2015] Any individual, group or organization that can affect, be affected by or perceive itself to be affected by a decision or activity.

Stakeholder satisfaction

- [ISO/IEC 25022, 2016] Degree to which stakeholder needs are satisfied when a product or system is used in a specified context of use.
  - Note 1: Users of a product or system are one type of stakeholder, so user satisfaction is one type of stakeholder satisfaction.

Usability

- [ISO 9241-210, 2019] Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use
  - The specified users, goals and context of use refer to the particular combination of users, goals and context of use for which usability is being considered.
  - The word usability is also used as a qualifier to refer to the design knowledge, competencies, activities and design attributes that contribute to usability, such as usability expertise, usability professional, usability engineering, usability method, usability evaluation, usability heuristic.
- [ISO/IEC 25010, 2011a] Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

User

- [ISO/IEC/IEEE 15288, 2015] Individual or group that interacts with a system or benefits from a system during its utilization.

User controllability

- Involved individual's possibility of avoiding harm in the situation that is putting him/her at risk

User error protection

- [ISO/IEC 25010, 2011a] Degree to which a product or system protects users against making errors

User interface aesthetics

- [ISO/IEC 25010, 2011a] Degree to which a user interface enables pleasing and satisfying interaction for the user.

Chapter K

System Engineering

K.1. Introduction

After the review of different definitions of AI-based systems, we aim to describe their distinctive aspects. In the following items, we summarise our findings. This synthetic view helps to differentiate AI-based from typical systems.

Convergence intelligence - systems engineering. The term intelligence is found and has been used in a variety of papers (including many research) from decades. Documents like [Rockwell Anyoha, 2017][1] and [OECD - AIGO, 2019] show that the term adopts different interpretations across history. All of them reflect the aim of reproducing human features ranging from perception of our physical environment up to execution of complex cognitive tasks (e.g. , gaming, mathematical reasoning) or even handling extra sensory capacities, like feelings, and emotions, or even meta-intelligent processes like intuition (perception of truth independent of reasoning). On the other side, systems engineering is nowadays an interdisciplinary area covering both, engineering and engineering management, dedicated to solve how to design, integrate, and manage complex systems in terms of associated life cycles. According to NASA, *Systems engineering is a methodical, disciplined approach for the design, realisation, technical management, operations, and retirement of a system.* [NASA, 2016]. ISO/IEC 15288 considers systems engineering as a process defining *the interdisciplinary tasks, which are required throughout a system's life cycle to transform customer needs, requirements, and constraints into a system solution* [ISO/IEC/IEEE, 2015]. The engineering of AI-based systems can be seen as the convergence of both areas: foundations for intelligence and engineering to develop systems integrating it. Such convergence imposes major challenges given (1) the complexity of reproducing human features and integrating them within a new generation of systems, and that (2) typical development cycles and engineering know-how do not suffice to accomplish the development of such systems.

[1] http://sitn. hms. harvard. edu/flash/2017/history-artificial-intelligence/

Increased levels of autonomy. Among the traits associated to intelligence, autonomy is one of the most distinctive. Autonomy is mostly measured by the reasoning capabilities in terms of goals/missions, tasks accomplishment, and their performance. AI-based systems are then committed to ensure a certain level of autonomy. However, typical systems already exhibit certain level of autonomy, e. g. , aircraft auto-piloting systems. Irrespective of the techniques and technology applied, AI-based systems are categorised in terms of the degree of human intervention expected during tasks/missions execution, in particular in case of unexpected scenarios like wrong stimuli, misbehavior, failure, and conflicting inputs. Given that the maximum autonomy level usually excludes any human intervention, the system should fully accomplish tasks/missions as expected and properly dealing with unbearable risks. Whereas for a typical system design, the requirements aim to circumvent and integrate conditions for nominal operation, AI-based systems shall relax such constraint assuming that the system shall be able to cope even with unexpected conditions. **Broader range of uncertainty.** Typical systems are allocated with requirements (functional and non-functional) thus defining a domain of design. For those systems, humans are supposed to intervene and takeover in order to ensure system control specially during critical safety scenarios, out of nominal operation, e. g. , autopilot disengagement due to conflicting sensors' signals. In the case of AI-Systems, which are expected to face "unknown" situations with certain degree of autonomy, the human intervention is excluded what leads to a broader scope of the domain of design. By doing so, a broader range of uncertainty should be considered at design time.

Ideal of mimic human physical, biological and social features. Self-healing, self-protecting, discovering, maturing, socialising, playing, cooperating, etc. are among desirable features specific to humans. The ideal for an AI-based system of imitating or reproducing such features is still work in progress and, in many respects, far from accomplishment. Confiance. ai is indeed part of such effort. However, since the engineering of AI-based systems is, in many aspects, inspired by that ideal, it is worth mentioning because it helps to clarify the goals, their scope as well as the potential challenges to be overcome. More concretely, those high-level objectives become inputs for the specification of an engineering process adapted to accomplish the ideal. The term "artificial" can be easily related to the fact that systems are expected to mimic human features.

Non-deterministic nature of AI algorithms. A wide variety of algorithms and techniques have been documented in the literature, e. g. Support Vector Machines (SVM), Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Gaussian Mixture Models (GMM). In addition, AI also comprises other non-machine learning techniques and methods like Genetic Algorithms, Fuzzy Logic or even Expert Systems. Irrespective of the category selected, a certain degree of non-determinism is incorporated by AI, for instance, the application of certain heuristics is well known in Genetic Algorithms. Non-determinism is a distinctive feature of AI-based systems which is particularly useful during solution searching, specially for exploring complex (n-dimensional, infinity, heterogeneous) problem spaces. However, non-determinism also raises some questions specially regarding our understanding on how and why AI algorithms find local or global solutions (i. e. , their explainability) and how to design an intelligence able to safely behave even in the case it's unable to find a solution.

Engineering process dependent on AI technology. Typical system development cycles rely on phases that aim to enrich design at system, software or hardware level. In the case of AI-based systems, the subsystems or components implementing ML/DL (Machine Learning/Deep Learning) modules are based upon parameters which may require to be set during design, implementation and validation phases. For those subsystems and components, a learning phase should be introduced whereas for other typical (non-ML-based) subsystems and components, the learning phase is non-existing. At least, an additional iteration phase is expected in order to conduct such training and tuning.

Engineering process dependent on knowledge bases. The learning phases previously referred strongly depend upon target objectives and external knowledge bases (KBs) necessary to accomplish them. For many complex AI-based systems, an iteration on design parameters may be necessary after a validation campaign, e. g. , to adjust detection ranges and improve accuracy. Detailed requirements cannot be elicited before knowing the effectiveness of knowledge bases, ML/DL techniques, and parameters choices.

Systems engineering concerns inherited by AI-Systems. The range of concerns currently treated in systems engineering is quite vast. In particular we can mention safety, security and, more recently, legal and ethical concerns like privacy and trust. For many emergent domains, e. g. , IoT (Internet of Things), C-ITS (Communicating Intelligent Transport Systems), etc. referred concerns are still challenges to be addressed and for which solutions are still to come. So far, the risks associated to those concerns should also be faced by AI-based systems. Consequently, their design should also integrate features to ensure a suitable response in alignment with their autonomy. At least two paradigms can be mentioned to accomplish such goal. The first approach consists in deploying typical approaches, techniques and processes to ensure correct incorporation of concerns (e.g. , MBSE, safety and security methods, formal methods). In the second approach, it is foreseen the application of AI techniques in order to address the related concerns.

Re-assessment of human and machine responsibilities. Given the increased levels of autonomy combined with a broader scope of uncertainty, it is reasonable to consider the likely occurrence of undesirable scenarios involving individuals harms, lives losses or other impacts like economical and environmental. To our knowledge, current legal frameworks and regulations shall still evolve in order to provide a basis for responsibilities assessment in case of unwanted events involving AI-based systems. It is expected that their deployment will motivate an evolution of legal orders and frameworks leading to upgrades of key concepts (like for instance *human error*) and to a reconfiguration of human's and machine's roles in society.

K.2. System Engineering Taxonomy

A posteriori-provability

- [Mamalet et al., 2021] The desired property is verified on the model after training. This approach may also rely on some assumptions on the ML algorithm (e.g. the architecture,

the size of the network, the activation function type for a NN...), but these assumptions depends on the problem

A priori-provability or by-design provability

- [Mamalet et al., 2021] The desired property is mathematically transferable as a design constraint to the ML algorithm. Then, to prove the property, it is necessary to demonstrate the validity of this transfer (i.e., if the design constraint is satisfied then the property holds on the model) and to demonstrate compliance with the design constraint.

Acceptance Criteria

- [ISO/IEC/IEEE 24765, 2017]
  - Criteria that a system or component must satisfy in order to be accepted by a user, customer, or other authorized entity.
  - A set of conditions that is required to be met before deliverables are accepted

Acceptance test

- [ISO/IEC/IEEE 24765, 2017] Test of a system or functional unit usually performed by the purchaser on his premises after installation with the participation of the vendor to ensure that the contractual requirements are met.

Accuracy

- [ISO/IEC/IEEE 24765, 2017]
  1. A qualitative assessment of correctness, or freedom from error.
  2. A quantitative measure of the magnitude of error.
  3. Within the quality management system, accuracy is an assessment of correctness.
- [ISO 17572, 2015] Measure of closeness of results of observations, computations, or estimates to the true values or the values accepted as being true

Anomaly

- [SAE AS6983, 2019] Data which is outside the ML Model ODD

Anomaly - misclassified samples

- [Corbière et al., 2019, Geifman and El-Yaniv, 2017] Objects that are likely to be misclassified and that fall near the decision boundary where the classifier is uncertain. Such problems are known as the problem of classification with reject option.

Anomaly - novelty detection

- [Schölkopf et al., 2001] Test points that could be new observations, i.e., the equivalent of an outlier for the test data. An example would be a new rare breed of dogs for a dataset of dog breeds. Such points are particularly interesting in active learning where once identified they are annotated and added to the training set, further improving the current classifier.

Anomaly - outlier detection

- [Rousseeuw and Driessen, 1999] Data points from the training set that are far from the others, i.e., an unusual or noisy training sample. An example would be a highly blurry picture of a pedestrian in a dataset for pedestrian detection or a picture of cat within a dataset of dogs.

Anomaly - out-of-distribution (OOD) detection

- [Liang et al., 2017, Malinin and Gales, 2018, Meinke and Hein, 2019, Ahmed and Courville, 2020] Objects that are drawn from a distribution different from the training distribution. In deep learning, we can distinguish two types of OOD:
  - low-level: different pixel statistics due to a mismatch between training and testing environments, e.g., a perception model for a vehicle trained for a country and tested on another one, a model trained on day-time images and tested on night-time images.
  - high-level (semantic): unknown objects or entities in a familiar environment, e.g., electric scooters for a perception model trained before scooters populated the streets

Assessment

- [CENELEC EN 50126, 2011] The undertaking of an investigation in order to arrive at a judgment, based on evidence, of the suitability of a product.
- [ISO/IEC 21827, 2008] Verification of a product, system or service against a standard using the corresponding assessment method to establish compliance and determine the assurance.

Asset

- [ISO/IEC 21827, 2008] Anything that has value to an organization.

Assurance

- [ISO/TS 21089, 2018] Development, documentation, testing, procedural and operational activities carried out to ensure a system's services do in fact provide the claimed level of function, performance and usability

Assurance Case

- [Rushby, 2015] An assurance case provides an argument to justify certain claims about a system, based on evidence concerning both the system and the environment in which it operates. The claims can be about any system property, such as reliability or security, and thereby generalize the previously established notion of a safety case, where the claim is always about safety.
- [ISO/IEC/IEEE 15026-1, 2019] Reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)
- [Mansourov and Campara, 2010] An assurance case is a structured argument, supported by evidence, intended to justify that a system is acceptably assured relative to a concern (such as safety or security) in the intended operating environment.

Audit

- [CENELEC EN 50126, 2011] A systematic and independent examination to determine whether the procedures specific to the requirements of a product comply with the planned arrangements, are implemented effectively and are suitable to achieve the specified objectives.
- [ISO/IEC 27000, 2018] A systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled.
  - Note 1: An audit can be an internal audit (first party) or an external audit (second party or third party), and it can be a combined audit (combining two or more disciplines).
  - Note 2: An internal audit is conducted by the organization itself, or by an external party on its behalf.

Auditability

- [Mamalet et al., 2021] The extent to which an independent examination of the development and verification process of the system can be performed

Availability

- [EN 50129, 2018] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.
- [ISO/IEC 27000, 2018] Property of being accessible and usable on demand by an authorized entity
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component is operational and accessible when required for use

Business-critical system

- A system whose failure may result in very high costs for the business using that system

Capacity

- [ISO/IEC 25010, 2011a] Degree to which the maximum limits of the product or system, parameter meet requirements.

Certification

- [ISO/IEC/IEEE 24765, 2017]
  1. Third-party attestation related to products, processes, systems, or persons.
  2. A written guarantee that a system or component complies with its specified requirements and is acceptable for operational use.
  3. Formal demonstration that a system or component complies with its specified requirements and is acceptable for operational use.
  4. Process of confirming that a system or component complies with its specified requirements and is acceptable for operational use.

Certification Credit

- [RTCA DO-297, 2005] Acceptance by the certification authority that a process, product or demonstration satisfies a certification requirement.

Compliance

- [CENELEC EN 50126, 2011] A demonstration that a characteristic or property of a product satisfies the stated requirements.

Confidentiality

- [ISO/IEC 25010, 2011a] Degree to which the prototype ensures that data are accessible only to those authorized to have access.

Configuration Management

- [CENELEC EN 50126, 2011] A discipline applying technical and administrative direction and surveillance to identify and document the functional and physical characteristics of a configuration item, control change to those characteristics, record and report change processing and implementation status and verify compliance with specified requirements.
- [RTCA DO-297, 2005] A discipline applying technical and administrative direction and surveillance to (a) identify and record the functional and physical characteristics of a configuration item, (b) control changes to those characteristics, and (c) record and report change control processing and implementation status.

Criticality

- [ISO/IEC/IEEE 24765, 2017] Degree of impact that a requirement, module, error, fault, failure, or other item has on the development or operation of a system.

Criticality Level

- [ANSI/UL 4600, 2020] Level categorizing the risk associated with an unmitigated hazard.

Dependability

- [Avizienis et al., 2004] The ability to deliver service that can justifiably be trusted. It entails:
  - Availability: readiness for correct service;
  - Reliability: continuity of correct service;
  - Safety: absence of catastrophic consequences on the user(s) and the environment;
  - Confidentiality: absence of unauthorized disclosure of information;
  - Integrity: absence of improper system alterations;
  - Maintainability: ability to undergo modifications, and repairs.
  - Security: the concurrent existence of availability for authorized users only, confidentiality, and integrity (with improper meaning unauthorized here)
- [High-Level Expert Group on Artificial Intelligence, 2019] Ability to deliver services that can justifiably be trusted.
- [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required
  - Note 1: Dependability includes availability, reliability, recoverability, maintainability, and maintenance support performance, and, in some cases, other characteristics such as durability, safety and security.
  - Note 2: Dependability is used as a collective term for the time-related quality characteristics of an item.
  - [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required

Design (verb)

- [ISO/IEC/IEEE 15288, 2015] The process to define the architecture, system elements, interfaces, and other characteristics of a system or system element.

Dynamic Assurance Case

- [ASAADI et al., 2020] Generic framework to provide justified confidence in the capabilities of autonomous systems embedding machine learning-based components.

Error

- [Avižienis et al., 2004] An error is defined as the part of a system s total state that may lead to a failure

Evidence

- [ISO/TS 21089, 2018] Everything that is used to determine or demonstrate the truth of an assertion

Failure

- [ISO 26262-1, 2018] Elimination of the ability of an element, to perform a function as required.

Fault tolerance

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component operates as intended despite the presence of hardware or software faults

Freedom from risk

- A characteristic that is defined as the degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment.

Functional failure

- [ISO 26262-1, 2018] Absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E [Eletrical / Electronic] systems.

Global Robustness

- [Mamalet et al., 2021] Ability of the system to perform the intended function in the presence of abnormal or unknown inputs

Governance

- [ISO/IEC 38500, 2015] System of directing and controlling.

Harm

- [ISO GUIDE 51, 2014] Injury or damage to the health of people or damage to property or the environment.

Hazard

- [ISO GUIDE 51, 2014] Potential source of harm.

High-risk AI System

- [European Commission, 2021] AI systems that create a high risk to the health and safety or fundamental rights of natural persons.

Incremental Acceptance

- [RTCA DO-297, 2005] A process for obtaining credit toward approval and certification by accepting or finding that an IMA module, application, and/or off-aircraft IMA system complies with specific requirements. Credit granted for individual tasks contributes to the overall certification goal. N.B. very domain specific.

Indicator

- [ISO/IEC 27000, 2018] Measure that provides an estimate or evaluation.

Innocuity

- [Holloway, 2019] Any part of the implementation that is not required by the defined intended behavior has no unacceptable impact.

Installability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency in which a product or system can be successfully installed and/or uninstalled in a specified environment.

Integrity

- [ISO/IEC 27000, 2018] Property of protecting the accuracy and completeness of assets.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
- [ISO/IEC TR 24028, 202O] An AI systems respect of sound moral and ethical principles or the assurance that information will not be manipulated in a malicious way by the AI system.

Integrity level

- [ISO/IEC/IEEE 24765, 2017]
  - Value representing project-unique characteristic, such as complexity, criticality, risk, safety level, security level, desired performance, and reliability, that define the importance of the system, software, or hardware to the user.
  - Degree to which software complies or must comply with a set of stakeholder-selected software and/or software-based system characteristics defined to reflect the importance of the software to its stakeholders.
  - Symbolic value representing a degree of compliance within an integrity level scheme.

  - Claim of a system, product, or element that includes limitations on a property's values, the claim's scope of applicability, and the allowable uncertainty regarding the claim's achievement.
  - Required degree of confidence that the system-of-interest meets the associated integrity level claim.

Intended behavior

- [ISO/PAS 21448, 2019] Specified behavior of the intended functionality including interaction with items.

Intended functionality

- [ISO/PAS 21448, 2019] behavior specified for a system.

Intent

- [Holloway, 2019] The defined intended behavior is correct and complete with respect to the desired behavior.

Interpretability

- [Arrieta et al., 2020] The ability to explain or to provide the meaning in understandable terms to a human.
- [Mamalet et al., 2021] Relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a human being. It is a basic requirement for an explanation.

Interpretable model

- An interpretable model should provide users with a description of what a stimulus, such as a datapoint or model output, means in context.

Learning Assurance

- [EASA, 2021] The learning assurance building block is intended to cover the paradigm shift from programming to learning, as the existing development assurance methods are not adapted to cover learning processes specific to AI/ML.

Level of risk

- [ISO/IEC 27000, 2018] Magnitude of a risk expressed in terms of the combination of consequences and their likelihood.

Life cycle

- [ISO/IEC/IEEE 15288, 2015] The evolution of a system, product, service, project or other human-made entity from conception through retirement. A life cycle can be described using an abstract functional model that represents the conceptualization of a need for the system, its realization, utilization, evolution and disposal.
- [ISO/IEC DIS 22989, 2021a] Evolution of a system, product, service, project or other human-made entity, from conception through retirement.

Local Robustness

- [Mamalet et al., 2021] The extent to which the system provides equivalent responses for similar inputs.

Maintainability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements
- The ability to identify and fix a fault within a software component is what the maintainability characteristic addresses. In other software quality models this characteristic is referenced as supportability. Maintainability is impacted by code readability or complexity as well as modularization. Anything that helps with identifying the cause of a fault and then fixing the fault is the concern of maintainability. Also the ability to verify (or test) a system, i.e. testability, is one of the subcharacteristics of maintainability.
- [Mamalet et al., 2021] Ability of extending/improving a given system while maintaining its compliance with the unchanged requirements..

Maturity

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component meets needs for reliability under normal operation.

Measurability

- [ISO/IEC TS 5723, 2022] Ability to assess an attribute of an entity against a metric Note 1: The word "measurable" is the adjective form of measurability.

Measure

- [ISO/IEC 25024, 2015] Variable to which a value is assigned as the result of measurement.
  - Note 1: The term measures is used to refer collectively to base measures, derived measures, and indicators.

Measure of Effectiveness

- [of Defense, 2020, Haskins et al., 2006] The operational measures of success that are closely related to the achievement of the mission or operational objective being evaluated, in the intended operational environment under a specified set of conditions; i.e., how well the solution achieves the intended purpose.

Measure of Performance

- [of Defense, 2020, Haskins et al., 2006] The measures that characterize physical or functional attributes relating to the system operation, measured or estimated under specified testing and/or operational environment conditions.

Measurement

- [ISO/IEC 25024, 2015] Set of operations having the object of determining a value of a measure

Measurement function

- [ISO/IEC 25024, 2015] Algorithm or calculation performed to combine two or more quality measure elements

Mission-critical system

- A system that is essential to the survival of a business or organization.

Misuse

- [ISO/PAS 21448, 2019] Usage of the system by a human in a way not intended by the manufacturer of the system.

ML Robustness

- [SAE AS6983, 2019] The capacity of an ML model to preserve its expected / intended performance under well-characterized abnormalities or deviations to its inputs and operating conditions outside its operational design domain (ODD)

ML Stability

- [SAE AS6983, 2019] The capacity of an ML model to preserve its expected / intended performance under well-characterized and bounded perturbations to its inputs and operating conditions within its operational design domain (ODD)

Monitor

- [Leucker and Schallhart, 2009] A monitor is a device that reads a finite trace and yield a certain verdict.

Non-overridable

- [EASA, 2021] Human has no capability to override the AI-based system s operations.

Non-repudiation

- [ISO/IEC 25010, 2011a] Degree to which actions or events can be proven to have taken place, so that the events or actions cannot be repudiated later.

Performance efficiency

- [ISO/IEC 25010, 2011a] Performance relative to the amount of resources used under stated conditions

Provability

- [Mamalet et al., 2021] The extent to which a set of properties on this algorithm can be guaranteed mathematically.

Quality

- [ISO/TS 13972, 2015] Degree to which all the properties and characteristics of a product, process or service satisfy the requirements which ensue from the purpose for which that product, process or service is to be used.

Quality Assurance

- [ISO/IEC/IEEE 15288, 2015] Part of quality management focused on providing confidence that quality requirements will be fulfilled.
- [Daniels et al., 2002, Haskins et al., 2006] Set of activities throughout the entire project life cycle necessary to provide adequate confidence that a product or service conforms to stakeholder requirements or that a process adheres to established methodology.
- Potential synonyms are: Assurance, Product Assurance, Development Assurance, Design Assurance.
- Related notions:
  - Safety Assurance (Quality Assurance in the scope of safety)
  - Quality Control (Inspection contributing to Quality Assurance)

Quality characteristic

- [ISO/IEC/IEEE 15288, 2015] Inherent characteristic of a product, process, or system related to a requirement.

Quality characteristic

- [ISO/IEC 25023, 2015] Category of quality attributes that bears on software product or system quality

Quality in use

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, satisfaction, and freedom from risk in specific contexts of use
  - Note 1: The quality in use of a software product or system can be measured and evaluated by the effect of the target system or software products when used by users of the implemented system or during field testing or prototype testing.
  - Note 2: When quality in use is specified, it relates to specified users meeting their needs to achieve specified goals with effectiveness, efficiency, satisfaction, and freedom from risk in specified contexts of use.

Quality Management

- [ISO/IEC/IEEE 15288, 2015] Coordinated activities to direct and control an organization with regard to quality.

Quality measure

- [ISO/IEC 25024, 2015] Measure that is defined as a measurement function of two or more values of quality measure elements

Quality measure element

- [ISO/IEC 25024, 2015] Measure defined in terms of a property and the measurement method for quantifying it, including optionally the transformation by mathematical function

Quality model

- [ISO/IEC 25024, 2015] Defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality

Recoverability

- [ISO/IEC 25010, 2011a] Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.

Reliability

- [ISO/IEC 27000, 2018] Property of consistent intended behavior and results.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.
- [ISO/IEC TS 5723, 2022] Ability of an item to perform as required, without failure, for a given time interval, under given conditions
- [SAE J3016, 2018] The probability that an item will perform a required function under specified conditions, without failure, for a specified period
- Ability of an item to perform as required, without failure, for a given time interval, under given condition
- [EASA, 2021] The probability that an item will perform a required function under specified conditions, without failure, for a specified period of time

Replaceability

- [ISO/IEC 25010, 2011a] Degree to which a product can replace another specified software product for the same purpose in the same environment.

Requirement

- [ISO/IEC/IEEE 15288, 2015] Statement that translates or expresses a need and its associated constraints and conditions.

Resilience

- [Mamalet et al., 2021] Ability for a system to continue to operate while an error or a fault has occurred
- [High-Level Expert Group on Artificial Intelligence, 2019] Robustness when facing changes.
- [ISO/IEC TS 5723, 2022] Capability of a system to maintain its functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Responsibility

- [ISO/IEC 38500, 2015] Obligation to act and take decisions to achieve required outcomes.

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Reusability

- [ISO/IEC 25010, 2011a] Degree to which an asset can be used in more than one system, or in building other assets

Robustness

- [ISO/IEC DIS 22989, 2021a, ISO/IEC TR 24029-1, 2021] Ability of a system to maintain its level of performance under a variety of circumstances.
- [Mamalet et al., 2021] (Global) Ability of the system to perform the intended function in the presence of abnormal or unknown inputs / (Local) The extent to which the system provides equivalent responses for similar inputs.
- [EASA, 2021] For an input varying in a region of the state space, the system is producing the same outputs.
- [Gehr et al., 2018] Local robustness (or robustness, for short) requires that all samples in the neighbourhood of a given input are classified with the same label.

Runtime monitor (runtime verification)

- [Cassar et al., 2017] Runtime Monitoring is a lightweight and dynamic verification technique that involves observing the internal operations of a software system and/or its interactions with other external entities, with the aim of determining whether the system satisfies or violates a correctness specification.

Scenario

- [ISO/PAS 21448, 2019] Description of the temporal development between several scenes in a sequence of scenes.

Scene

- [ISO/PAS 21448, 2019] Snapshot of the environment including the scenery, dynamic elements, and all actor and observer self representations, and the relationships between those entities.

Situation

- [ISO/PAS 21448, 2019] Selection of an appropriate behavior pattern at a particular point of time.

Specifiability

- [Mamalet et al., 2021] The extent to which the system can be correctly and completely described through a list of requirements (such as stakeholder requirements, "black box" requirements, or "white box" requirements).
- Requirement: an identifiable element of a function specification that can be validated and against which an implementation can be verified.

Testability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.

Time behavior

- [ISO/IEC 25010, 2011a] Degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements

Traceability

- [IEEE 610.12-1990, 2002]
  1. The degree to which a relationship can be established between, two or more products of the development process, especially products having a predecessor, successor, or master-subordinate relationship to one another; for example, the degree to which the requirements and design of a given software component match.
  2. The degree to which each element in a software development product establishes its reason for existing; for example, the degree to which each element in a bubble chart references the requirement that it satisfies.
- [EASA, 2020] An association between artifacts, such as between process outputs or between an output and its originating process
- [EASA, 2021] The ability to track the journey of a data input through all stages of sampling, labelling, processing and decision-making

Transparency

- [ISO/IEC DIS 22989, 2021b]
  - <organization> Property of an organization that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner.

  - <system> Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability.
- [Arrieta et al., 2020] A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models are divided into three categories: simulatable models, decomposable models and algorithmically transparent models.
- [Brundage et al., 2020] Making information about the characteristics of an AI devel oper s operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.
- [IEEE 7000, 2021] Transparency means that information provided about a system is meaningful, useful, accessible, comprehensive, and truthful.
  1. Meaningful means that information about a system should be relevant for users concern or user control.
  2. Usefulness of information implies that consumers can act upon it and make choices easily, acting upon the information provided to them.
  3. Accessible means that it is possible to easily obtain and retrieve the relevant information in a machine-readable or other way whether through state-of-the-art electronic channels or via constrained devices or constrained networks.
  4. Comprehensive means that information about a system should be easy to read and understand for ordinary people and to require any expert knowledge.
  5. Truthful means that information about a system accurately reflects a system s or system landscape s activities, such as data processing and data sharing practices. The information should be up to date and written in plain language that is clear and direct. It should not mislead users in any way, hide information, or give half-truth about practices.
- Related and Opposing values
  - Related values: Openness, cleanliness, explicability, explainability, access to data, auditability
  - Opposing values: Privacy, bribery, corruption
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability
- [ISO/IEC DIS 22989, 2021b] Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [Brundage et al., 2020] Making information about the characteristics of an AI developer s operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.

Triggering event

- [ISO/PAS 21448, 2019] Specific conditions of a driving scenario that serve as an initiator for a subsequent system reaction possibly leading to a hazardous event.

Trust

- [ISO/IEC 25010, 2011b] Degree to which a user or other stakeholder has confidence that a product or system will behave as intended.
- [IEEE 7000, 2021] Trust in a system can be granted as a result of a system s demonstrated competence, benevolence, honesty and predictable behavior.
  1. System competence is a matter of system dependability; that is system security, reliability, and safety.
  2. Dependability can be signaled to users through some evidence or frame, such as quality seals or certification, publicly stated guarantees, and warranties.
  3. System benevolence is embedded in human-computer interaction, which can be of emotional, responsive, and respectful manner.
  4. System honesty can be signaled by a system through ts way of being transparent.
  5. System predictability is fostered by embedding standardized forms of interaction (signaling situation normality) and making a system sustainable and easy-to-use
- Related and Opposing values
  - Related values: Predictability, dependability, veracity.
  - Opposing values: Control.

Trustworthiness

- [ISO/IEC TR 24028, 2020] Ability to meet stakeholders expectations in a verifiable way.
  - Note 1: Depending on the context or sector, and also on the specific product or service, data, and technology used, different characteristics apply and need verification to ensure stakeholders expectations are met.
  - Note 2: Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.
  - Note 3: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.
- [High-Level Expert Group on Artificial Intelligence, 2019] Trustworthy AI, as defined by the High level expert group on AI from the European Union is
  1. lawful, i.e. complying with all applicable laws and regulations
  2. ethical, i.e. ensuring adherence to ethical principles and values
  3. robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm

Uncertainty

- [ISO/IEC 2382, 2015] Condition appearing when a value cannot be determined during consultation, or a fact or a rule in the knowledge base remains in doubt.

Validation

- [Aerospace, 2010] The determination that the requirements for a product are correct and complete. (Are we building the right aircraft / system / function / item?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled (the right system was built).

Value

- [ISO 10303-1, 2021] Belief(s) an organization adheres to and the standards that it seeks to observe.

Verifiability

- [Mamalet et al., 2021] Ability to evaluate an implementation of requirements to determine that they have been met (adapted from ARP4754A)

Verification

- [Aerospace, 2010] The evaluation of an implementation of requirements to determine that they have been met. (Did we build the aircraft / system/ function / item right?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled (the system was built right).

Verifiable

- [ISO/IEC/IEEE 15288, 2015] Can be checked for correctness by a person or tool

Vulnerability

- [ISO/IEC 27000, 2018] Weakness of an asset or control that can be exploited by one or more threats.
- [ISO GUIDE 73, 2009] Intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence.

Chapter L

Trustworthiness AI

L.1. Introduction

L.2. Trustworthiness AI Taxonomy

A posteriori-provability

- [Mamalet et al., 2021] The desired property is verified on the model after training. This approach may also rely on some assumptions on the ML algorithm (e.g. the architecture, the size of the network, the activation function type for a NN...), but these assumptions depends on the problem

A priori-provability or by-design provability

- [Mamalet et al., 2021] The desired property is mathematically transferable as a design constraint to the ML algorithm. Then, to prove the property, it is necessary to demonstrate the validity of this transfer (i.e., if the design constraint is satisfied then the property holds on the model) and to demonstrate compliance with the design constraint.

Acceptance Criteria

- [ISO/IEC/IEEE 24765, 2017]
  - Criteria that a system or component must satisfy in order to be accepted by a user, customer, or other authorized entity.
  - A set of conditions that is required to be met before deliverables are accepted

Acceptance test

- [ISO/IEC/IEEE 24765, 2017] Test of a system or functional unit usually performed by the purchaser on his premises after installation with the participation of the vendor to ensure

that the contractual requirements are met.

Accountability

- [ISO/IEC TR 24028, 2020] Property that ensures that the actions of an entity may be traced uniquely to that entity
- [ISO 7498-2, 1989] For systems, accountability is a property that ensures that actions of an entity can be traced uniquely to the entity.
- [ISO/IEC 25010, 2011a] Degree to which the actions of an entity can be traced uniquely to the entity.
- [EASA, 2021] Accountability refers to the idea that one is responsible for their action Ð and as a corollary their consequences Ð and must be able to explain their aims, motivations, and reasons.

Accuracy

- [ISO/IEC/IEEE 24765, 2017]
  1. A qualitative assessment of correctness, or freedom from error.
  2. A quantitative measure of the magnitude of error.
  3. Within the quality management system, accuracy is an assessment of correctness.
- [ISO 17572, 2015] Measure of closeness of results of observations, computations, or estimates to the true values or the values accepted as being true

Adaptability

- [ISO/IEC 25010, 2011a] Degree to which a product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments.

AI explainability

- [EASA, 2021] The AI explainability building block deals with the capability to provide the human with understandable and relevant information on how an AI/ML application is coming to its results.

AI safety risk

- [EASA, 2021] The AI safety risk mitigation building block considers that we may not always be able to open the AI black box to the extent required and that the safety risk may need to be addressed to deal with the inherent uncertainty of AI.

Assessment

- [CENELEC EN 50126, 2011] The undertaking of an investigation in order to arrive at a judgment, based on evidence, of the suitability of a product.

- [ISO/IEC 21827, 2008] Verification of a product, system or service against a standard using the corresponding assessment method to establish compliance and determine the assurance.

Assurance

- [ISO/TS 21089, 2018] Development, documentation, testing, procedural and operational activities carried out to ensure a system's services do in fact provide the claimed level of function, performance and usability

Assurance Case

- [Rushby, 2015] An assurance case provides an argument to justify certain claims about a system, based on evidence concerning both the system and the environment in which it operates. The claims can be about any system property, such as reliability or security, and thereby generalize the previously established notion of a safety case, where the claim is always about safety.
- [ISO/IEC/IEEE 15026-1, 2019] Reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)
- [Mansourov and Campara, 2010] An assurance case is a structured argument, supported by evidence, intended to justify that a system is acceptably assured relative to a concern (such as safety or security) in the intended operating environment.

Audit

- [CENELEC EN 50126, 2011] A systematic and independent examination to determine whether the procedures specific to the requirements of a product comply with the planned arrangements, are implemented effectively and are suitable to achieve the specified objectives.
- [ISO/IEC 27000, 2018] A systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled.
  - Note 1: An audit can be an internal audit (first party) or an external audit (second party or third party), and it can be a combined audit (combining two or more disciplines).
  - Note 2: An internal audit is conducted by the organization itself, or by an external party on its behalf.

Auditability

- [Mamalet et al., 2021] The extent to which an independent examination of the development and verification process of the system can be performed

Authenticity

- [ISO/IEC 25010, 2011a] Degree to which the identity of a subject or resource can be proved to be the one claimed assets

Availability

- [EN 50129, 2018] The ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval assuming that the required external resources are provided.
- [ISO/IEC 27000, 2018] Property of being accessible and usable on demand by an authorized entity
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component is operational and accessible when required for use

Bias

- [ISO/IEC TR 29119-11, 2020] Measure of the distance between the predicted value provided by the ML model and a desired fair prediction.
- [ISO/IEC TR 24028, 2020] Favouritism towards some things, people or groups over others.

Black box

- [ISO/IEC/IEEE 24765, 2017]
  1. A system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or implementation are unknown or irrelevant.
  2. Pertaining to an approach that treats a system or component whose inputs, outputs, and general function are known but whose contents or implementation are unknown or irrelevant.

Capacity

- [ISO/IEC 25010, 2011a] Degree to which the maximum limits of the product or system, parameter meet requirements.

Certification

- [ISO/IEC/IEEE 24765, 2017]
  1. Third-party attestation related to products, processes, systems, or persons.
  2. A written guarantee that a system or component complies with its specified requirements and is acceptable for operational use.
  3. Formal demonstration that a system or component complies with its specified requirements and is acceptable for operational use.

4. Process of confirming that a system or component complies with its specified requirements and is acceptable for operational use.

Certification Credit

- [RTCA DO-297, 2005] Acceptance by the certification authority that a process, product or demonstration satisfies a certification requirement.

Comprehensibility

- [Arrieta et al., 2020] When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. This notion of model comprehensibility stems from the postulates of Michalski, which stated that the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single chunks of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity.
- [Arrieta et al., 2020] when conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.

Compliance

- [CENELEC EN 50126, 2011] A demonstration that a characteristic or property of a product satisfies the stated requirements.

Completeness of explainability

- [Mamalet et al., 2021] Relates to the capability to describe a phenomenon in such a way that this description can be used to reach a given goal.

Component

- [Szyperski et al., 2002] A basic building-block for systems with well-defined interfaces, behavior and explicit context dependencies only. A component can be deployed independently. That is, it implements a clear function. A component can be composed with other components into systems, sub-system or new components. A component can exist in the form of software or hardware or a combination of both.

Confidentiality

- [ISO/IEC 25010, 2011a] Degree to which the prototype ensures that data are accessible only to those authorized to have access.

Configuration Management

- [CENELEC EN 50126, 2011] A discipline applying technical and administrative direction and surveillance to identify and document the functional and physical characteristics of a configuration item, control change to those characteristics, record and report change processing and implementation status and verify compliance with specified requirements.
- [RTCA DO-297, 2005] A discipline applying technical and administrative direction and surveillance to (a) identify and record the functional and physical characteristics of a configuration item, (b) control changes to those characteristics, and (c) record and report change control processing and implementation status.

Completeness of explainability

- [Mamalet et al., 2021] Relates to the capability to describe a phenomenon in such a way that this description can be used to reach a given goal.

Comprehensibility

- [Arrieta et al., 2020] When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. This notion of model comprehensibility stems from the postulates of Michalski, which stated that the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single chunks of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity.
- [Arrieta et al., 2020] when conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.

Context completeness

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used with the required levels of effectiveness, efficiency, satisfaction, and freedom from risk in each of the specified contexts of use.
  - Note 1: Context completeness is a subcharacteristic of context coverage.

Context coverage

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used with effectiveness, efficiency, satisfaction, and freedom from risk in both specified contexts of use and in contexts beyond those initially explicitly identified
  - Note 1: Context of use is relevant to both quality in use and some product quality (sub)characteristics (where it is referred to as specified conditions).

Context of use

- [ISO/IEC 25022, 2016] Users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a system, product or service is used

Control

- [IEEE 7000, 2021] Having control of a machine means having
  1. Cognitive control in terms of being informed about what is going on in the computing environment,
  2. Decisional control in terms of having choices over what is going on in one s networked environment,
  3. Behavioral control in terms of receiving feedback on one s actions/choices taken.

  Related and Opposing values
  - Related values: Human responsibility, governance, usability, portability,logic, sense of accomplishment, moderation.
  - Opposing values: Trust, accountability to stakeholders; imagination, reminding, obedience.

Controllability

- [ISO 26262-1, 2018] Ability to avoid a specified harm or damage through the timely reactions of the persons involved, possibly with support from external measures.
- [ISO 26262-1, 2011] Ability to avoid a specified harm or damage through the timely reactions of the persons involved, possibly with support from external measures

Correctness

- [ISO/IEC/IEEE 24765, 2017]
  1. The degree to which a system or component is free from faults in its specification, design, and implementation.
  2. The degree to which software, documentation, or other items meet specified requirements.
  3. The degree to which software, documentation, or other items meet user needs and expectations, whether specified or not
- [Holloway, 2019] The implementation is correct with respect to its defined intended behavior, under foreseeable operating conditions.

- [ISO/IEC/IEEE 24765, 2017] Degree to which a system or component is free from faults in its specification, design, and implementation.

Criticality

- [ISO/IEC/IEEE 24765, 2017] Degree of impact that a requirement, module, error, fault, failure, or other item has on the development or operation of a system.

Criticality Level

- [ANSI/UL 4600, 2020] Level categorizing the risk associated with an unmitigated hazard.

Dependability

- [Avizienis et al., 2004] The ability to deliver service that can justifiably be trusted. It entails Availability: readiness for correct service; Reliability: continuity of correct service; Safety: absence of catastrophic consequences on the user(s) and the environment; Confidentiality: absence of unauthorized disclosure of information; Integrity: absence of improper system alterations; Maintainability: ability to undergo modifications, and repairs. Security: the concurrent existence of availability for authorized users only, confidentiality, and integrity (with improper meaning unauthorized here)
- [High-Level Expert Group on Artificial Intelligence, 2019] Ability to deliver services that can justifiably be trusted.
- [ISO/IEC/IEEE 15026-1, 2019] Ability to perform as and when required

Effectiveness

- [ISO 9000, 2015] Extent to which planned activities are realized and planned results achieved.
- [ISO 9241-210, 2019] Accuracy and completeness with which users achieve specified goals

Efficiency

- [ISO 9000, 2015] Relationship between the results achieved and the resources used.
- Relationship between the results achieved and the resources used. Resources expended in relation to the accuracy and completeness with which users achieve goals

Error

- [Avižienis et al., 2004] An error is defined as the part of a system s total state that may lead to a failure

Evidence

- [ISO/TS 21089, 2018] Everything that is used to determine or demonstrate the truth of an assertion

Ethical

- [Cambridge Dictionary, 2020] Morally right (or, alternatively: relating to beliefs about what is morally right and wrong).

Explainability

- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.
- [Phillips et al., 2020] Explanation: Systems deliver accompanying evidence or reason(s) for all outputs.
  - Meaningful: Systems provide explanations that are understandable to individual users.
  - Explanation Accuracy: The explanation correctly reflects the system s process for generating the output.
  - Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.
- [Arrieta et al., 2020] Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.
- [Mamalet et al., 2021] The extent to which the behavior of a model can be made understandable to humans
- [ISO/IEC DIS 22989, 2021a] Property of an AI system to express important factors influencing the AI system results in a way that humans can understand.

Explainable Artificial Intelligence (XAI)

- [Arrieta et al., 2020] An explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Explainable model

- An explanation of a model result is a description of how a model s outcome came to be.

Explanation facility

- [ISO/IEC 2382, 2015] Component of a knowledge-based system that explains how solutions were derived and justifies the steps used in reaching them.

Failure

- [ISO 26262-1, 2018] Elimination of the ability of an element, to perform a function as required.

Fairness

- [Stevenson, 2015] Impartial and just treatment or behavior without favoritism or discrimination.
- [High-Level Expert Group on Artificial Intelligence, 2019] Fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as substantive fairness. But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated
- [IEEE 7000, 2021] Fairness has the attributes of systematic discrimination with an absence of bias in reaching reasonable judgments and allowing opportunities.
  Related and Opposing values
  - Related Values: Responsible position on conflicts of interest, tolerance, justice, balance, equality (legal, gender, minority)
  - Opposing values: Bias, suspicion, discrimination, arbitrariness

Fault tolerance

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component operates as intended despite the presence of hardware or software faults

Freedom from risk

- A characteristic that is defined as the degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment.

Functional appropriateness

- [ISO/IEC 25010, 2011a] Degree to which the functions facilitate the accomplishment of specified tasks and objectives

Functional completeness

- [ISO/IEC 25010, 2011a] Degree to which the set of functions covers all the specified tasks and user objectives

Functional correctness

- [ISO/IEC 25010, 2011a] Degree to which a product or system provides the correct results with the needed degree of precision

Functional failure

- [ISO 26262-1, 2018] Absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E [Eletrical / Electronic] systems.

Functional suitability

- [ISO/IEC 25010, 2011a] Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

Global Robustness

- [Mamalet et al., 2021] Ability of the system to perform the intended function in the presence of abnormal or unknown inputs

Harm

- [ISO GUIDE 51, 2014] Injury or damage to the health of people or damage to property or the environment.

Hazard

- [ISO GUIDE 51, 2014] Potential source of harm.

High-risk AI System

- [European Commission, 2021] AI systems that create a high risk to the health and safety or fundamental rights of natural persons.

Human Cognitive Bias

- [ISO/IEC TR 24027, 2021] Bias that occurs when humans are processing and interpreting information Note 1: human cognitive bias influences judgement and decision-making.

Inclusiveness

- [IEEE 7000, 2021] Inclusiveness in a system means that it is accessible to differently abled users, unbiased in its decisions, and fair to the broadest range of characteristics (especially human characteristics) it may encounter.
  Related and Opposing values
  - Related values: Participation, partnership, solidarity, interdependence,compatibility, accessibility, diversity
  - Opposing values: Control, bias, detachment

Incremental Acceptance

- [RTCA DO-297, 2005] A process for obtaining credit toward approval and certification by accepting or finding that an IMA module, application, and/or off-aircraft IMA system complies with specific requirements. Credit granted for individual tasks contributes to the overall certification goal. N.B. very domain specific.

Indicator

- [ISO/IEC 27000, 2018] Measure that provides an estimate or evaluation.

Installability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency in which a product or system can be successfully installed and/or uninstalled in a specified environment.

Integrity

- [ISO/IEC 27000, 2018] Property of protecting the accuracy and completeness of assets.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component prevents unauthorized access to, or modification of, computer programs or data.
- [ISO/IEC TR 24028, 202O] An AI system's respect of sound moral and ethical principles or the assurance that information will not be manipulated in a malicious way by the AI system.
- [ISO/IEC/IEEE 24765, 2010]
    1. Value representing project-unique characteristic, such as complexity, criticality, risk, safety level, security level, desired performance, and reliability, that define the importance of the system, software, or hardware to the user.
    2. Degree to which software complies or must comply with a set of stakeholder-selected software and/or software-based system characteristics defined to reflect the importance of the software to its stakeholders.
    3. Symbolic value representing a degree of compliance within an integrity level scheme.
    4. Claim of a system, product, or element that includes limitations on a property's values, the claim's scope of applicability, and the allowable uncertainty regarding the claim's achievement.
    5. Required degree of confidence that the system-of-interest meets the associated integrity level claim.

Integrity level

- [ISO/IEC/IEEE 24765, 2017]
    - Value representing project-unique characteristic, such as complexity, criticality, risk, safety level, security level, desired performance, and reliability, that define the importance of the system, software, or hardware to the user.

  - Degree to which software complies or must comply with a set of stakeholder-selected software and/or software-based system characteristics defined to reflect the importance of the software to its stakeholders.
  - Symbolic value representing a degree of compliance within an integrity level scheme.
  - Claim of a system, product, or element that includes limitations on a property's values, the claim's scope of applicability, and the allowable uncertainty regarding the claim's achievement.
  - Required degree of confidence that the system-of-interest meets the associated integrity level claim.

Intended behavior

- [ISO/PAS 21448, 2019] Specified behavior of the intended functionality including interaction with items.

Intended functionality

- [ISO/PAS 21448, 2019] behavior specified for a system.

Intent

- [Holloway, 2019] The defined intended behavior is correct and complete with respect to the desired behavior.

Interpretability

- [Arrieta et al., 2020] The ability to explain or to provide the meaning in understandable terms to a human.
- [Mamalet et al., 2021] Relates to the capability of an element representation (an object, a relation, a property...) to be associated with the mental model of a human being. It is a basic requirement for an explanation.

Interpretable model

- An interpretable model should provide users with a description of what a stimulus, such as a datapoint or model output, means in context.

Learnability

- [ISO/IEC 25010, 2011a] Degree to which a product or system enables the user to learn how to use it with effectiveness, efficiency in emergency situations.

Learning Assurance

- [EASA, 2021] The learning assurance building block is intended to cover the paradigm shift from programming to learning, as the existing development assurance methods are not adapted to cover learning processes specific to AI/ML.

Level of risk

- [ISO/IEC 27000, 2018] Magnitude of a risk expressed in terms of the combination of consequences and their likelihood.

Life cycle

- [ISO/IEC/IEEE 15288, 2015] The evolution of a system, product, service, project or other human-made entity from conception through retirement. A life cycle can be described using an abstract functional model that represents the conceptualization of a need for the system, its realization, utilization, evolution and disposal.
- [ISO/IEC DIS 22989, 2021a] Evolution of a system, product, service, project or other human-made entity, from conception through retirement.

Local Robustness

- [Mamalet et al., 2021] The extent to which the system provides equivalent responses for similar inputs.

Local Robustness

- [Mamalet et al., 2021] The extent to which the system provides equivalent responses for similar inputs.

Maintainability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which a product or system can be modified to improve it, correct it or adapt it to changes in environment, and in requirements
- The ability to identify and fix a fault within a software component is what the maintainability characteristic addresses. In other software quality models this characteristic is referenced as supportability. Maintainability is impacted by code readability or complexity as well as modularization. Anything that helps with identifying the cause of a fault and then fixing the fault is the concern of maintainability. Also the ability to verify (or test) a system, i.e. testability, is one of the sub-characteristics of maintainability.
- [Mamalet et al., 2021] Ability of extending/improving a given system while maintaining its compliance with the unchanged requirements..

Maturity

- [ISO/IEC 25010, 2011a] Degree to which a system, product or component meets needs for reliability under normal operation.

Measurability

- [ISO/IEC TS 5723, 2022] Ability to assess an attribute of an entity against a metric
- Note 1: The word "measurable" is the adjective form of measurability.

Measure

- [ISO/IEC 25024, 2015] Variable to which a value is assigned as the result of measurement Note 1: The term measures is used to refer collectively to base measures, derived measures, and indicators.

Measure of Effectiveness

- [of Defense, 2020, Haskins et al., 2006] The operational measures of success that are closely related to the achievement of the mission or operational objective being evaluated, in the intended operational environment under a specified set of conditions; i.e., how well the solution achieves the intended purpose.

Measure of Performance

- [of Defense, 2020, Haskins et al., 2006] The measures that characterize physical or functional attributes relating to the system operation, measured or estimated under specified testing and/or operational environment conditions.

Measurement

- [ISO/IEC 25024, 2015] Set of operations having the object of determining a value of a measure

Measurement function

- [ISO/IEC 25024, 2015] Algorithm or calculation performed to combine two or more quality measure elements

Metric

- In mathematics, metric may refer to one of two related, but distinct concepts:
  - A function which measures distance between two points in a metric space
  - A metric tensor, in differential geometry, which allows defining lengths of curves, angles, and distances in a manifold

- In Engineering and business: The word metric is often used to mean a descriptive statistic, indicator, or figure of merit used to describe or measure something quantitatively, including:
  - Performance indicator, a measure of an organization's activities and performance
  - Software metric, a measure of some property of a piece of software or its specifications
  - Reuse metrics, a quantitative indicator of an attribute for software reuse and reusability
  - Search engine optimization metrics, indicators of a website's organic search potential

Misuse

- [ISO/PAS 21448, 2019] Usage of the system by a human in a way not intended by the manufacturer of the system.

Modifiability

- [ISO/IEC 25010, 2011a] Degree to which a product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.

Modularity

- [ISO/IEC 25010, 2011a] Degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.

Monitor

- [Leucker and Schallhart, 2009] A monitor is a device that reads a finite trace and yield a certain verdict.

ML Robustness

- [SAE AS6983, 2019] The capacity of an ML model to preserve its expected / intended performance under well-characterized abnormalities or deviations to its inputs and operating conditions outside its operational design domain (ODD)

ML Stability

- [SAE AS6983, 2019] The capacity of an ML model to preserve its expected / intended performance under well-characterized and bounded perturbations to its inputs and operating conditions within its operational design domain (ODD)

Modifiability

- [ISO/IEC 25010, 2011a] Degree to which a product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.

Modularity

- [ISO/IEC 25010, 2011a] Degree to which a system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.

Monitor

- [Leucker and Schallhart, 2009] A monitor is a device that reads a finite trace and yield a certain verdict.

Non-overridable

- [EASA, 2021] Human has no capability to override the AI-based systemÕs operations.

Non-repudiation

- [ISO/IEC 25010, 2011a] Degree to which actions or events can be proven to have taken place, so that the events or actions cannot be repudiated later.

Performance efficiency

- [ISO/IEC 25010, 2011a] Performance relative to the amount of resources used under stated conditions

Portability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another

Precision

- [ISO/IEC TR 29119-11, 2020] Performance metric used to evaluate a classifier, which measures the proportion of predicted positives that were correct.
- [ISO/DIS 5725-1, 2020] Closeness of agreement between independent test results obtained under stipulated conditions
  - Note 1: Precision depends only on the distribution of random errors and does not relate to the true value or the specified value.

  - Note 2: The measure of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results. Less precision is reflected by a larger standard deviation.
  - Note 3: Quantitative measures of precision depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme conditions.

Precision of Explanability

- [Mamalet et al., 2021] Indicates how much details must be provided to the human to let her/him execute mentally the inference in a right way with respect to her/his goal. For instance, there is no need to know the laws of general relativity or quantum mechanics to predict the trajectory of a ball.

Predictability

- [EASA, 2021] The degree to which a correct forecast of a system's state can be made quantitatively.

Privacy

- [ISO/IEC 2382, 2015] Freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual.
- [IEEE 7000, 2021] Privacy means that the collection, processing, and dissemination of personal information is done in such a way as to maintain the information self-determination of a data subject.
- Related and Opposing values
  - Related values: Respect for confidentiality, intimacy, anonymity.
  - Opposing values: Transparency,inclusiveness, alerting.

Process

- [ISO 9000, 2015] Set of interrelated or interacting activities that use inputs to deliver an ntended result.

Provability

- [Mamalet et al., 2021] The extent to which a set of properties on this algorithm can be guaranteed mathematically.

Quality

- [ISO/TS 13972, 2015] Degree to which all the properties and characteristics of a product, process or service satisfy the requirements which ensue from the purpose for which that

product, process or service is to be used.

Quality Assurance

- [ISO/IEC/IEEE 15288, 2015] Part of quality management focused on providing confidence that quality requirements will be fulfilled.
- [Daniels et al., 2002, Haskins et al., 2006] Set of activities throughout the entire project life cycle necessary to provide adequate confidence that a product or service conforms to stakeholder requirements or that a process adheres to established methodology.
- Potential synonyms are: Assurance, Product Assurance, Development Assurance, Design Assurance.
- Related notions:
  - Safety Assurance (Quality Assurance in the scope of safety)
  - Quality Control (Inspection contributing to Quality Assurance)

Quality characteristic

- [ISO/IEC/IEEE 15288, 2015] Inherent characteristic of a product, process, or system related to a requirement.
- [ISO/IEC 25023, 2015] Category of quality attributes that bears on software product or system quality

Quality in use

- [ISO/IEC 25022, 2016] Degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, satisfaction, and freedom from risk in specific contexts of use
  - Note 1: The quality in use of a software product or system can be measured and evaluated by the effect of the target system or software products when used by users of the implemented system or during field testing or prototype testing.
  - Note 2: When quality in use is specified, it relates to specified users meeting their needs to achieve specified goals with effectiveness, efficiency, satisfaction, and freedom from risk in specified contexts of use.

Quality Management

- [ISO/IEC/IEEE 15288, 2015] Coordinated activities to direct and control an organization with regard to quality.

Quality measure

- [ISO/IEC 25024, 2015] Measure that is defined as a measurement function of two or more values of quality measure elements

Quality measure element

- [ISO/IEC 25024, 2015] Measure defined in terms of a property and the measurement method for quantifying it, including optionally the transformation by mathematical function

Quality model

- [ISO/IEC 25024, 2015] Defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality

Recoverability

- [ISO/IEC 25010, 2011a] Degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.

Reliability

- [ISO/IEC 27000, 2018] Property of consistent intended behavior and results.
- [ISO/IEC 25010, 2011a] Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.
- [ISO/IEC TS 5723, 2022] Ability of an item to perform as required, without failure, for a given time interval, under given conditions
- [SAE J3016, 2018] The probability that an item will perform a required function under specified conditions, without failure, for a specified period
- Ability of an item to perform as required, without failure, for a given time interval, under given condition
- [EASA, 2021] The probability that an item will perform a required function under specified conditions, without failure, for a specified period of time

Replaceability

- [ISO/IEC 25010, 2011a] Degree to which a product can replace another specified software product for the same purpose in the same environment.

Requirement

- [ISO/IEC/IEEE 15288, 2015] Statement that translates or expresses a need and its associated constraints and conditions.

Resilience

- [Mamalet et al., 2021] Ability for a system to continue to operate while an error or a fault has occurred

- [High-Level Expert Group on Artificial Intelligence, 2019] Robustness when facing changes.
- [ISO/IEC TS 5723, 2022] Capability of a system to maintain its functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Respect

- [IEEE 7000, 2021] Respect in human-machine interaction implies that a machine is perceived as attentive and responsive. a/ Attentiveness implies that the machine is perceived as replying in a reasonable amount of time and respecting user privacy. b/ Responsiveness implies that the machine is perceived as applying appropriate criteria in its decisions and made explicit to the user and that it is perceived as acting fairly and politely
  Related and Opposing values
  - Related values: Politeness, courtesy, respect for environment and natural habitat, respect for information and confidentiality, respect for norms, reputation.
  - Opposing values: Self-esteem, maleficence.

Responsibility

- [ISO/IEC 38500, 2015] Obligation to act and take decisions to achieve required outcomes.

Responsibility

- [ISO/IEC 38500, 2015] Obligation to act and take decisions to achieve required outcomes.

Resource utilization

- [ISO/IEC 25010, 2011a] Degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements.

Restricted Operation Domain

- [Colwell et al., 2018] The specific conditions under which a given driving automation system or feature thereof is currently able to function, including, but not limited to, driving modes.

Reusability

- [ISO/IEC 25010, 2011a] Degree to which an asset can be used in more than one system, or in building other assets

Risk

- [CENELEC EN 50126, 2011] The probable rate of occurrence of a hazard causing harm and the degree of severity of the harm.
- [DEFSTAN 00-56(PT1)/7, 2017] Combination of the likelihood of harm and the severity of that harm or long term damage to health.
- [ANSI/UL 4600, 2020] A combination of the probability of occurrence of a loss event and the severity of that loss event.
- [ISO/IEC 27000, 2018] Effect of uncertainty on objectives.
  - Note 1: An effect is a deviation from the expected Ñ positive or negative.
  - Note 2: Uncertainty is the state, even partial, of deficiency of information related to, understanding or knowledge of, an event, its consequence, or likelihood.
  - Note 3: Risk is often characterized by reference to potential events (as defined in ISO-Guide 73:2009, 3.5.1.3) and consequences (as defined in ISO-Guide 73:2009, 3.6.1.3), or a combination of these.
  - Note 4: Risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood (as defined in ISO-Guide 73:2009, 3.6.1.1) of occurrence.
  - Note 5: In the context of information security management systems, information security risks can be expressed as effect of uncertainty on information security objectives.
  - Note 6: Information security risk is associated with the potential that threats will exploit vulnerabilities of an information asset or group of information assets and thereby cause harm to an organization.

Risk Assessment

- [ISO/IEC 27000, 2018] Overall process of risk identification, risk analysis and risk evaluation.

Risk management process

- [ISO GUIDE 73, 2009] Systematic application of management policies, procedures and practices to the activities of communicating, consulting, establishing the context, and identifying, analyzing, evaluating, treating, monitoring and reviewing risk.

Risk mitigation

- [ISO 22300, 2021] Lessening or minimizing of the adverse impacts of a hazardous event.

Robustness

- [ISO/IEC DIS 22989, 2021a, ISO/IEC TR 24029-1, 2021] Ability of a system to maintain its level of performance under a variety of circumstances.

- [Mamalet et al., 2021] (Global) Ability of the system to perform the intended function in the presence of abnormal or unknown inputs / (Local) The extent to which the system provides equivalent responses for similar inputs.
- [EASA, 2021] For an input varying in a region of the state space, the system is producing the same outputs.
- [Gehr et al., 2018] Local robustness (or robustness, for short) requires that all samples in the neighborhood of a given input are classified with the same label.

Runtime monitor (runtime verification)

- [Cassar et al., 2017] Runtime Monitoring is a lightweight and dynamic verification technique that involves observing the internal operations of a software system and/or its interactions with other external entities, with the aim of determining whether the system satisfies or violates a correctness specification.

Satisfaction

- [ISO/IEC 25022, 2016] Degree to which user needs are satisfied when a product or system is used in a specified context of use
  - Note 1: For a user who does not directly interact with the product or system, only purpose accomplishment and trust are relevant.
  - Note 2: Satisfaction is the user s response to interaction with the product or system, and includes attitudes towards use of the product.
  - Note 3: Users include: primary users who interact with the system to achieve the primary goals, secondary users who provide support, and indirect users who receive output, but do not interact with the system.
  - Note 4: In this International Standard, user's needs include their desires and expectations associated with use of a product, system, or service. Exceeding desires and expectations is a means of significantly increasing satisfaction and improving the user experience.

Scenario

- [ISO/PAS 21448, 2019] Description of the temporal development between several scenes in a sequence of scenes.

Scene

- [ISO/PAS 21448, 2019] Snapshot of the environment including the scenery, dynamic elements, and all actor and observer self representations, and the relationships between those entities.

Security

- [ISO/IEC 25010, 2011b] Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.
- [ISO/IEC 23643, 2020] Resistance to intentional, unauthorized act(s) designed to cause harm or damage to a system

Semantic Data Accuracy

- [ISO/IEC 25012, 2008a] Closeness of the data values to a set of values defined in a domain considered semantically correct.

Situation

- [ISO/PAS 21448, 2019] Selection of an appropriate behavior pattern at a particular point of time.

Software Assurance

- [Hinchey et al., 2006] Software Assurance is the planned and systematic set of activities that ensures that software processes and products conform to requirements, standards and procedures

Specifiability

- [Mamalet et al., 2021] The extent to which the system can be correctly and completely described through a list of requirements (such as stakeholder requirements, "black box" requirements, or "white box" requirements).
- Requirement: an identifiable element of a function specification that can be validated and against which an implementation can be verified.

Stakeholder

- [ISO/IEC 38500, 2015] Any individual, group or organization that can affect, be affected by or perceive itself to be affected by a decision or activity.

Stakeholder satisfaction

- [ISO/IEC 25022, 2016] Degree to which stakeholder needs are satisfied when a product or system is used in a specified context of use.
  - Note 1: Users of a product or system are one type of stakeholder, so user satisfaction is one type of stakeholder satisfaction.

Sustainability

- [IEEE 7000, 2021] Related and Opposing values - Related values: Respect for environment and natural habitat, efficiency, maintainability, operability, supportability, reliability, durability, resilience, forgiveness, robustness, redundancy, reusability, re-configurability, simplicity, economy, renewability. - Opposing values: Cost (extravagance), wastefulness, poverty, consumption

System-Dependent Data Quality

- [ISO/IEC 25012, 2008a] Degree to which data quality is reached and preserved within a computer system when data is used under specified conditions.

Test Dataset

- [Confiance.ai, 2021b] A Dataset that is only composed of Observations that will be used only for evaluating the ML Model performance under operational configuration. Test Dataset must be fully independent of Training Dataset and Validation Dataset. No Observation from this set can be part of these two Datasets.

Testability

- [ISO/IEC 25010, 2011a] Degree of effectiveness and efficiency with which test criteria can be established for a system, product or component and tests can be performed to determine whether those criteria have been met.

Time behavior

- [ISO/IEC 25010, 2011a] Degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements

Traceability

- [IEEE 610.12-1990, 2002] (1) The degree to which a relationship can be established between, two or more products of the development process, especially products having a predecessor, successor, or master-subordinate relationship to one another; for example, the degree to which the requirements and design of a given software component match. (2) The degree to which each element in a software development product establishes its reason for existing; for example, the degree to which each element in a bubble chart references the requirement that it satisfies.
- [EASA, 2020] An association between artifacts, such as between process outputs or between an output and its originating process
- [EASA, 2021] The ability to track the journey of a data input through all stages of sampling, labeling, processing and decision-making

Transparency

- [ISO/IEC DIS 22989, 2021b]
  - <organization> Property of an organization that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner.
  - <system> Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability.
- [Arrieta et al., 2020] A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models are divided into three categories: simulatable models, decomposable models and algorithmically transparent models.
- [Brundage et al., 2020] Making information about the characteristics of an AI devel oper s operations or their AI systems available to actors both inside and outside the organization. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.
- [IEEE 7000, 2021] Transparency means that information provided about a system is meaningful, useful, accessible, comprehensive, and truthful.
  1. Meaningful means that information about a system should be relevant for users concern or user control.
  2. Usefulness of information implies that consumers can act upon it and make choices easily, acting upon the information provided to them.
  3. Accessible means that it is possible to easily obtain and retrieve the relevant information in a machine-readable or other way whether through state-of-the-art electronic channels or via constrained devices or constrained networks.
  4. Comprehensive means that information about a system should be easy to read and understand for ordinary people and to require any expert knowledge.
  5. Truthful means that information about a system accurately reflects a system s or system landscape s activities, such as data processing and data sharing practices. The information should be up to date and written in plain language that is clear and direct. It should not mislead users in any way, hide information, or give half-truth about practices.
- Related and Opposing values
  - Related values: Openness, cleanliness, explicability, explainability, access to data, auditability
  - Opposing values: Privacy, bribery, corruption
- [ISO/IEC 27036-3, 2013] Property of a system or process to imply openness and accountability
- [ISO/IEC DIS 22989, 2021b] Property of a system that appropriate information about the system is communicated to relevant stakeholders.
- [Brundage et al., 2020] Making information about the characteristics of an AI developer s operations or their AI systems available to actors both inside and outside the organiza-

tion. In recent years, transparency has emerged as a key theme in work on the societal implications of AI.

Trueness

- [ISO/DIS 5725-1, 2020] Closeness of agreement between the expectation of test results and a true value
  - Note 1: The measure of trueness is usually expressed in terms of bias.
  - Note 2: Trueness is sometimes referred to as accuracy of the mean. This usage is not recommended.
  - Note 3: In practice, the accepted reference value is substituted for the true value.

Trust

- [ISO/IEC 25010, 2011b] Degree to which a user or other stakeholder has confidence that a product or system will behave as intended.
- [IEEE 7000, 2021] Trust in a system can be granted as a result of a systems demonstrated competence, benevolence, honesty and predictable behavior.
  1. System competence is a matter of system dependability; that is system security, reliability, and safety.
  2. Dependability can be signaled to users through some evidence or frame, such as quality seals or certification, publicly stated guarantees, and warranties.
  3. System benevolence is embedded in human-computer interaction, which can be of motional, responsive, and respectful manner.
  4. System honesty can be signaled by a system through ts way of being transparent.
  5. System predictability is fostered by embedding standardized forms of interaction (signaling situation normality) and making a system sustainable and easy-to-use
- Related and Opposing values
  - Related values: Predictability, dependability, veracity.
  - Opposing values: Control.

Trustworthiness

- [ISO/IEC TR 24028, 2020] Ability to meet stakeholders expectations in a verifiable way.
  - Note 1: Depending on the context or sector, and also on the specific product or service, data, and technology used, different characteristics apply and need verification to ensure stakeholders expectations are met.
  - Note 2: Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.
  - Note 3: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.

- [High-Level Expert Group on Artificial Intelligence, 2019] Trustworthy AI, as defined by the High level expert group on AI from the European Union is
  1. lawful, i.e. complying with all applicable laws and regulations
  2. ethical, i.e. ensuring adherence to ethical principles and values
  3. robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm

Uncertainty

- [ISO/IEC 2382, 2015] Condition appearing when a value cannot be determined during consultation, or a fact or a rule in the knowledge base remains in doubt.

Usability

- [ISO 9241-210, 2019] Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use
  - The specified users, goals and context of use refer to the particular combination of users, goals and context of use for which usability is being considered.
  - The word usability is also used as a qualifier to refer to the design knowledge, competencies, activities and design attributes that contribute to usability, such as usability expertise, usability professional, usability engineering, usability method, usability evaluation, usability heuristic.
- [ISO/IEC 25010, 2011a] Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

User

- [ISO/IEC/IEEE 15288, 2015] Individual or group that interacts with a system or benefits from a system during its utilization.

User controllability

- Involved individual's possibility of avoiding harm in the situation that is putting him/her at risk

User error protection

- [ISO/IEC 25010, 2011a] Degree to which a product or system protects users against making errors

User interface aesthetics

- [ISO/IEC 25010, 2011a] Degree to which a user interface enables pleasing and satisfying interaction for the user.

Validation

- [Aerospace, 2010] The determination that the requirements for a product are correct and complete. (Are we building the right aircraft / system / function / item?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled (the right system was built).

Value

- [ISO 10303-1, 2021] Belief(s) an organization adheres to and the standards that it seeks to observe.

Verifiability

- [Mamalet et al., 2021] Ability to evaluate an implementation of requirements to determine that they have been met (adapted from ARP4754A)

Verification

- [Aerospace, 2010] The evaluation of an implementation of requirements to determine that they have been met. (Did we build the aircraft / system/ function / item right?)
- [ISO/IEC/IEEE 15288, 2015] Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled (the system was built right).

Verifiable

- [ISO/IEC/IEEE 15288, 2015] Can be checked for correctness by a person or tool

Vulnerability

- [ISO/IEC 27000, 2018] Weakness of an asset or control that can be exploited by one or more threats.
- [ISO GUIDE 73, 2009] Intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence.

Chapter M

AI Job Families

AI scientist

An AI Scientist is someone who designs and creates artificial intelligence, improves the efficiency of artificial intelligence systems as well as the application of AI. As research scientist, AI Scientist will have to be expert in multiple disciplines of AI, including computational statistics, data-driven AI (as machine learning), knowledge-based AI, as well as hybrid AI and embedded or distributed AI. The work of an AI scientist requires them to be well-versed in math, statistics, and programming.

AI engineer

AI engineer focuses on developing the tools, systems, and processes that enable artificial intelligence to be applied in the real world. Any application where machines mimic human functions, such as solving problems and learning, can be considered artificial intelligence.

Algorithm engineer

Algorithm Engineers design, implement and improve algorithms to optimize the operation of (critical) systems. Algorithm Engineers are highly skilled specialists who use math and science to solve both simple and complex everyday problems. Algorithm engineers do more than write new algorithms when required. They are also responsible for testing their algorithms against expectations, gauging technology and data efficiency to inform results.

(Big) Data analyst

A data analyst is the one who collects, organizes and analyzes large sets of data (known as Big Data) to discover patterns and some other useful information. Data mining and Data auditing are a must have skills to become a Data Analyst.

Business intelligence developer

As a business intelligence developer, his/her primary goal will be to analyze complex datasets and to identify business as well as market trends in such a way as to boost his/her organization's revenue. Using cloud-based data platforms, his/her task will be design, model, as well as maintaining complex data.

Data engineer

Data Engineers build and optimize the systems that allow data scientists and analysts to perform their work. The Data Engineers establish the foundation that the data analysts and scientists build upon. Data Engineers are also responsible for constructing data pipelines and often have to use complex tools and techniques to handle data at scale.

Data scientist

A data scientist will be dealing with extremely large and complex datasets. Data scientist applies his/her expertise in statistics and building machine learning models to make predictions and answer key business questions. He/she uses both machine learning as well as predictive analytics. He/she is also able to create algorithms that enable the gathering as well as the cleaning for such a huge amount of data, thereby preparing for it to be analyzed. However, a data scientist has more depth and expertise in his/her skills, and is also able to train and optimise machine learning models.

Data Algo and AI Discipline Manager

Data/Algo/AI Discipline Managers are responsible for inspiring and catalyzing the improvement of best practices within Artificial Intelligence, Algorithm, Data ensuring they are mastered effectively. They act as a privileged point of contact on complex subjects and take on an educational role so that their team can reach discipline excellence and deliver competitive solutions within a multidisciplinary environment.

Discipline Managers are managers-coaches serving discipline's engineers and offering a diverse and inclusive environment with Hungry Humble Aware values. They are the guardians of an engineering discipline excellence in a learning environment which contribute to highly competitive realizations (Quality, Attractiveness, Cost, Lead time).

Data Transformation rngineers

Data Transformation Engineers drive and steer the data transformation, adapting the Group data strategy, coordinating and supervising resources and efforts related to all data topics and recurrent activities. Their efforts are focused on efficiency gains and savings, while implementing the appropriate data foundations.

Intelligence developer

AI Engineer Build AI models from scratch and help product managers and stakeholders understand results.

Machine Learning engineer

Their tasks will involve his/her ability to apply predictive models and make efficient use of natural language processing as you deal with massive datasets. Their resume will be further strengthened if you have analytical skills, have experience in neural networks and deep learning, as well as possess necessary cloud applications.

Natural Language Processing engineer

Explore the connection between human language and computational systems; this includes working on projects like chatbots and virtual assistants.

M.0.1 Software engineer

Software engineers are computer science professionals who use engineering principles and programming languages to build software and run network control systems. Software engineers play an important role in making sure computers and mobile devices operate correctly. They bring a considerable amount of knowledge to roles in the areas of programming languages, software development and computer operating systems. They must also understand engineering principles as they relate to the creation of software applications and systems.

Software engineers are strategically minded individuals who tend to excel in left- and right-brained thinking (analytical as well as creative skills). They are usually instinctive problem solvers, able to use tools such as the principles of applied mathematics and computer science to design, develop and troubleshoot computer software. Ideally, software engineers should also be people that work well with others and are motivated to see a project through to the end.

All software engineers have the shared mission of solving digital problems with quality (debugged) software.

Alphabetical Index

Bibliography

[Ackoff, 1989] Ackoff, R. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1):3–9.

[Aerospace, 2010] Aerospace, S. (2010). Aerospace recommended practice arp4754 issued revised rev. a 1996-11 2010-12 superseding arp4754 (r) guidelines for.

[Ahmed and Courville, 2020] Ahmed, F. and Courville, A. (2020). Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3154–3162.

[Akhtar and Mian, 2018] Akhtar, N. and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430.

[Albers et al., 2009] Albers, S., Alt, H., and Näher, S. (2009). *Efficient algorithms: essays dedicated to Kurt Mehlhorn on the occasion of his 60th birthday*, volume 5760. Springer.

[ANSI/UL 4600, 2020] ANSI/UL 4600 (2020). Standard for Safety for the Evaluation of Autonomous Products.

[Arrieta et al., 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

[Artificial Intelligence Roadmap, 2020] Artificial Intelligence Roadmap (2020). A human-centric approach to ai in aviation. *European Aviation Safety Agency*.

[ASAADI et al., 2020] ASAADI, E., PETROFF, D., DENNEY, E., MENZIES, J., and PAI, G. (December 2020). Dynamic assurance cases: A pathway to trusted autonomy. *KBR Inc and NASA Ames Research center*.

[Ashmore and Madahar, 2019] Ashmore, R. and Madahar, B. (2019). Rethinking diversity in the context of autonomous systems. In *Engineering Safe Autonomy, 27th Safety-Critical Systems Symposium*, pages 175–192.

[Avižienis et al., 2004] Avižienis, A., Laprie, J.-C., and Randell, B. (2004). Dependability and its threats: a taxonomy. In *Building the Information Society*, pages 91–120. Springer.

[Avizienis et al., 2004] Avizienis, A., Laprie, J.-C., Randell, B., and Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE transactions on dependable and secure computing*, 1(1):11–33.

[Bak and Duggirala, 2017] Bak, S. and Duggirala, P. S. (2017). Simulation-equivalent reachability of large linear systems with inputs. In *International Conference on Computer Aided Verification*, pages 401–420. Springer.

[Barreno et al., 2006] Barreno, M., Nelson, B., Sears, R., Joseph, A. D., and Tygar, J. D. (2006). Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 16–25.

[Batini et al., 2009] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):1–52.

[Biggio and Roli, 2018] Biggio, B. and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331.

[Branco et al., 2008] Branco, M., Zaluska, E., De Roure, D., Salgado, P., Garonne, V., Lassnig, M., and Rocha, R. (2008). Managing very-large distributed datasets. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 775–792. Springer.

[Broniatowski et al., 2021] Broniatowski, D. A. et al. (2021). Psychological foundations of explainability and interpretability in artificial intelligence. *NIST, Tech. Rep*.

[Brundage et al., 2020] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al. (2020). Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

[Cai and Zhu, 2015] Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14.

[Cambridge Dictionary, 2020] Cambridge Dictionary (2020). Ethical. *Cambridge Dictionary*.

[Cassar et al., 2017] Cassar, I., Francalanza, A., Aceto, L., and Ingólfsdóttir, A. (2017). A Survey of Runtime Monitoring Instrumentation Techniques. *Electron. Proc. Theor. Comput. Sci.*, 254:15–28.

[CENELEC EN 50126, 2011] CENELEC EN 50126 (2011). Railway applications - The specification and demonstration of Reliability, Availability, Maintainability and Safety (RAMS).

[CENELEC EN 50128, 2020] CENELEC EN 50128 (2020). Railway applications - Communication, signalling and processing systems - Software for railway control and protection systems.

[Chakraborty et al., 2018] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

[Colwell et al., 2018] Colwell, I., Phan, B., Saleem, S., Salay, R., and Czarnecki, K. (2018). An automated vehicle safety concept based on runtime restriction of the operational design domain. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1910–1917.

[Confiance.ai, 2021a] Confiance.ai (2021a). EC2: Algorithm Engineering including Algorithm evaluation and KPIs State of the Art. Technical report, Confiance.ai Program.

[Confiance.ai, 2021b] Confiance.ai (2021b). EC5 L5.1.4.1 Taxonomy Data Definitions and Concepts. Technical report, Confiance.ai Program.

[Confiance.ai, 2021c] Confiance.ai (2021c). Project EC5: Data and knowledge engineering for trusted AI - Data exploration and qualification. Technical report, Confiance.ai Program.

[Corbière et al., 2019] Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. (2019). Addressing failure prediction by learning model confidence. *arXiv preprint arXiv:1910.04851*.

[Daneshjo, 2012] Daneshjo, N. (2012). Computers modeling and simulation. In *Advanced Materials Research*, volume 463, pages 1102–1105. Trans Tech Publ.

[Daniels et al., 2002] Daniels, S. E., Johnson, K., and Johnson, C. (2002). Quality glossary. *Quality Progress*, 35(7):43.

[DEFSTAN 00-56(PT1)/7, 2017] DEFSTAN 00-56(PT1)/7 (2017). Uk ministry of defence standards. safety management requirements for defence systems - part 1: Requirements.

[Demetrescu et al., 2004] Demetrescu, C., Finocchi, I., and Italiano, G. F. (2004). Algorithm engineering. In *Current Trends in Theoretical Computer Science: The Challenge of the New Century Vol 1: Algorithms and Complexity Vol 2: Formal Models and Semantics*, pages 83–104. World Scientific.

[EASA, 2020] EASA (2020). EASA Artificial Intelligence Roadmap 1.0 published. A human-centric approach to AI in aviation.

[EASA, 2021] EASA (2021). Concept paper first usable guidance for level 1 machine learning applications.

[ED-76A, 2015] ED-76A (2015). Standards for processing aeronautical data.

[EN 50129, 2018] EN 50129 (2018). Railway applications - Communication, signalling and processing systems - Safety-related electronic systems for signalling.

[EUROCAE ED 218, 2012] EUROCAE ED 218 (2012). *MODEL-BASED DEVELOPMENT AND VERIFICATION SUPPLEMENT TO ED-12C AND ED-109A*. European Organization for Civil Aviation Equipment (EUROCAE).

[European Commission, 2018] European Commission (2018). COMMUNICATION FROM THE COMMISSION Artificial Intelligence for Europe. COM/2018/237 final. Technical report, European Commission.

[European Commission, 2021] European Commission (2021). 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.

[Foggia et al., 2001] Foggia, P., Genna, R., and Vento, M. (2001). Symbolic vs. connectionist learning: an experimental comparison in a structured domain. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):176–195.

[Friedenthal et al., 2007] Friedenthal, S., Izumi, L., and Meilich, A. (2007). 9.2.2 Object-Oriented Systems Engineering Method (OOSEM) applied to Joint Force Projection (JFP), a Lockheed Martin Integrating Concept (LMIC). In *INCOSE International Symposium*, volume 17, pages 1471–1491. Wiley Online Library.

[Gehr et al., 2018] Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. (2018). Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

[Geifman and El-Yaniv, 2017] Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[Green et al., 2011] Green, B., Marotta, J., Petre, B., Lillestolen, K., Spencer, R., Gupta, N., O'Leary, D., Lee, J. D., Strasburger, J., Nordsieck, A., et al. (2011). Handbook for the selection and evaluation of microprocessors for airborne systems.

[Group et al., 2018] Group, A. C. W. et al. (2018). Goal structuring notation community standard (version 2).

[Gudivada et al., 2017] Gudivada, V., Apon, A., and Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1):1–20.

[Harnad, 1990] Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

[Haskins et al., 2006] Haskins, C., Forsberg, K., Krueger, M., Walden, D., and Hamelin, D. (2006). Systems engineering handbook. In *INCOSE*, volume 9, pages 13–16.

[High-Level Expert Group on Artificial Intelligence, 2019] High-Level Expert Group on Artificial Intelligence (2019). Assessment list for trustworthy artificial intelligence (altai). Technical report, European Commission.

[Hinchey et al., 2006] Hinchey, M. G., Pressburger, T., Markosian, L., and Feather, M. S. (2006). The nasa software research infusion initiative: successful technology transfer for software assurance. In *Proceedings of the 2006 international workshop on Software technology transfer in software engineering*, pages 43–48.

[HLEG, 2018] HLEG (2018). A definition of AI: main capabilities and scientific disciplines. Definition developed for the purpose of the deliverables of the High-Level Expert Group on AI.

[Hofer-Schmitz and Stojanović, 2020] Hofer-Schmitz, K. and Stojanović, B. (2020). Towards formal verification of iot protocols: A review. *Computer Networks*, 174:107233.

[Holloway, 2019] Holloway, C. M. (2019). Understanding the overarching properties.

[ICF, 2018] ICF (2018). JWS special issue on Knowledge Graphs/en.

[IEEE 610.12-1990, 2002] IEEE 610.12-1990 (2002). Ieee standard glossary of software engineering terminology.

[IEEE 7000, 2021] IEEE 7000 (2021). Model Process for Addressing Ethical Concerns During System Design.

[ISO 10303-1, 2021] ISO 10303-1 (2021). Industrial automation systems and integration — product data representation and exchange — part 1: Overview and fundamental principles.

[ISO 16269-4, 2010] ISO 16269-4 (2010). Statistical interpretation of data — Part 4: Detection and treatment of outliers.

[ISO 17572, 2015] ISO 17572 (2015). Intelligent transport systems (ITS) — Location referencing for geographic databases — Part 1: General requirements and conceptual model.

[ISO 22300, 2021] ISO 22300 (2021). Security and resilience — Vocabulary.

[ISO 26262-1, 2011] ISO 26262-1 (2011). Road vehicles — functional safety — part 1: Vocabulary.

[ISO 26262-1, 2018] ISO 26262-1 (2018). Road vehicles - Functional safety - Part 1: Vocabulary.

[ISO 7498-2, 1989] ISO 7498-2 (1989). Information processing systems — Open Systems Interconnection — Basic Reference Model — Part 2: Security Architecture.

[ISO 9000, 2015] ISO 9000 (2015). Quality management systems — Fundamentals and vocabulary.

[ISO 9241-210, 2019] ISO 9241-210 (2019). Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.

[ISO GUIDE 51, 2014] ISO GUIDE 51 (2014). Safety aspects — Guidelines for their inclusion in standards.

[ISO GUIDE 73, 2009] ISO GUIDE 73 (2009). Risk management — Vocabulary.

[ISO/DIS 5725-1, 2020] ISO/DIS 5725-1 (2020). Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions.

[ISO/IEC 11557, 1992] ISO/IEC 11557 (1992). Information technology.

[ISO/IEC 12207, 2017] ISO/IEC 12207 (2017). Systems and software engineering — software life cycle processes.

[ISO/IEC 20546, 2019] ISO/IEC 20546 (2019). Information technology — Big data — Overview and vocabulary.

[ISO/IEC 21827, 2008] ISO/IEC 21827 (2008). Information technology — Security techniques — Systems Security Engineering — Capability Maturity Model© (SSE-CMM©).

[ISO/IEC 23643, 2020] ISO/IEC 23643 (2020). Software and systems engineering — Capabilities of software safety and security verification tools.

[ISO/IEC 2382, 2015] ISO/IEC 2382 (2015). Information technology — vocabulary.

[ISO/IEC 2382-28, 1995] ISO/IEC 2382-28 (1995). Information technology — vocabulary — part 28: Artificial intelligence — basic concepts and expert systems.

[ISO/IEC 25010, 2011a] ISO/IEC 25010 (2011a). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models.

[ISO/IEC 25010, 2011b] ISO/IEC 25010 (2011b). *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*. ISO/IEC JTC 1/SC 7 Software and systems engineering.

[ISO/IEC 25012, 2008a] ISO/IEC 25012 (2008a). Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model.

[ISO/IEC 25012, 2008b] ISO/IEC 25012 (2008b). Software engineering — software product quality requirements and evaluation (square) — data quality model.

[ISO/IEC 25022, 2016] ISO/IEC 25022 (2016). Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — Measurement of quality in use.

[ISO/IEC 25023, 2015] ISO/IEC 25023 (2015). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of system and software product quality.

[ISO/IEC 25024, 2015] ISO/IEC 25024 (2015). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality.

[ISO/IEC 27000, 2018] ISO/IEC 27000 (2018). Information technology — Security techniques — Information security management systems — Overview and vocabulary.

[ISO/IEC 27036-3, 2013] ISO/IEC 27036-3 (2013). *Information technology — Security techniques — Information security for supplier relationships — Part 3: Guidelines for information and communication technology supply chain security*. ISO/IEC JTC 1/SC 27 Information security, cybersecurity and privacy protection.

[ISO/IEC 38500, 2015] ISO/IEC 38500 (2015). Information technology - governance of it for the organization.

[ISO/IEC DIS 22989, 2021a] ISO/IEC DIS 22989 (2021a). Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.

[ISO/IEC DIS 22989, 2021b] ISO/IEC DIS 22989 (2021b). *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*. ISO/IEC JTC 1/SC 42/WG 1 Foundational standards.

[ISO/IEC TR 24027, 2021] ISO/IEC TR 24027 (2021). *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. British Standards Institution - Information systems co-ordination.

[ISO/IEC TR 24028, 2020] ISO/IEC TR 24028 (2020). Information technology — artificial intelligence — overview of trustworthiness in artificial intelligence.

[ISO/IEC TR 24028, 202O] ISO/IEC TR 24028 (202O). *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*. British Standards Institution - Information systems co-ordination.

[ISO/IEC TR 24029-1, 2021] ISO/IEC TR 24029-1 (2021). Artificial intelligence (ai) — assessment of the robustness of neural networks — part 1: Overview.

[ISO/IEC TR 29119-11, 2020] ISO/IEC TR 29119-11 (2020). Software and systems engineering — software testing — part 11: Guidelines on the testing of ai-based systems.

[ISO/IEC TS 5723, 2022] ISO/IEC TS 5723 (2022). Trustworthiness — Vocabulary.

[ISO/IEC/IEEE, 2015] ISO/IEC/IEEE (2015). 15288:2015 Ingénierie des systèmes et du logiciel — Processus du cycle de vie du système.

[ISO/IEC/IEEE 15026-1, 2019] ISO/IEC/IEEE 15026-1 (2019). Systems and software engineering — Systems and software assurance — Part 1: Concepts and vocabulary.

[ISO/IEC/IEEE 15288, 2015] ISO/IEC/IEEE 15288 (2015). Systems and software engineering — System life cycle processes.

[ISO/IEC/IEEE 24765, 2010] ISO/IEC/IEEE 24765 (2010). Systems and software engineering — vocabulary.

[ISO/IEC/IEEE 24765, 2017] ISO/IEC/IEEE 24765 (2017). Systems and software engineering — vocabulary.

[ISO/PAS 21448, 2019] ISO/PAS 21448 (2019). Road vehicles — safety of the intended functionality.

[ISO/TR 23482-1, 2020] ISO/TR 23482-1 (2020). Robotics — application of iso 13482 — part 1: Safety-related test methods.

[ISO/TS 13972, 2015] ISO/TS 13972 (2015). Health informatics — Detailed clinical models, characteristics and processes.

[ISO/TS 21089, 2018] ISO/TS 21089 (2018). Health informatics — Trusted end-to-end information flows.

[JORF, 2018] JORF (2018). Jorf n°0285 du 9 décembre 2018.

[JRC, 2018] JRC (2018). Artificial intelligence: A european perspective. ec jrc flagship report on ai. EC JRC Flagship report on AI - ISBN 978-92-79-97217-1 ISSN 1831-9424 doi:10.2760/11251.

[Kasabov, 2012] Kasabov, N. (2012). Evolving spiking neural networks for spatio-and spectro-temporal pattern recognition. In *2012 6th IEEE International Conference Intelligent Systems*, pages 27–32.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2.

[Leucker and Schallhart, 2009] Leucker, M. and Schallhart, C. (2009). A brief account of runtime verification. *Journal of Logic and Algebraic Programming*, page 11.

[Liang et al., 2017] Liang, S., Li, Y., and Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

[Liu et al., 2018] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., and Leung, V. C. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, 6:12103–12117.

[Logility, 2021] Logility (2021). What's the difference between descriptive, predictive and prescriptive analytics?

[Malinin and Gales, 2018] Malinin, A. and Gales, M. (2018). Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*.

[Mamalet et al., 2021] Mamalet, F., Jenn, E., Flandrin, G., Delseny, H., Gabreau, C., et al. (2021). White Paper Machine Learning in Certified Systems. Research report, IRT Saint Exupéry ; ANITI.

[Mansourov and Campara, 2010] Mansourov, N. and Campara, D. (2010). *System assurance: beyond detecting vulnerabilities*. Elsevier.

[Meinke and Hein, 2019] Meinke, A. and Hein, M. (2019). Towards neural networks that provably know when they don't know. *arXiv preprint arXiv:1909.12180*.

[Müller-Hannemann and Schirra, 2001] Müller-Hannemann, M. and Schirra, S. (2001). *Algorithm Engineering*. Springer.

[NASA, 2016] NASA (2016). *NASA Systems Engineering Handbook*. National Aeronautics and Space Administration.

[NHTSA et al., 2016] NHTSA et al. (2016). *Federal automated vehicles policy: Accelerating the next revolution in roadway safety*. US Department of Transportation.

[Object Management Group, 2010] Object Management Group (2010). Mda foundation model. omg document number ormsc/2010-09-06. Technical report, Object Management Group.

[OECD - AIGO, 2019] OECD - AIGO (2019). Artificial Intelligence in Society.

[of Defense, 1996] of Defense, U. D. (1996). Dod modeling and simulation (m& s) verification, validation, and accreditation (vv& a). Technical report, OFFICE OF THE UNDER SECRETARY OF DEFENSE.

[of Defense, 2020] of Defense, U. D. (2020). Dod 5000.2 operation of the adaptive acquisition framework. Technical report, Office of the Under Secretary of Defense for Acquisition and Sustainment.

[Papernot et al., 2018] Papernot, N., McDaniel, P., Sinha, A., and Wellman, M. P. (2018). Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE.

[Phillips et al., 2020] Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., and Przybocki, M. A. (2020). *Four Principles of Explainable Artificial Intelligence*. NIST National Institute of Standards and Technology, U.S. Department of Commerce.

[Picard et al., 2020] Picard, S., Chapdelaine, C., Cappi, C., Gardes, L., Jenn, E., Lefevre, B., and Soumarmon, T. (2020). Ensuring dataset quality for machine learning certification. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 275–282. IEEE.

[Pipino et al., 2002] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218.

[Rockwell Anyoha, 2017] Rockwell Anyoha (2017). Can Machines Think?

[Rousseeuw and Driessen, 1999] Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

[RTCA DO-297, 2005] RTCA DO-297 (2005). Integrated modular avionics (ima) development guidance and certification considerations.

[Rushby, 2015] Rushby, J. (January 2015). On the interpretation of assurance case arguments. *SRI International*.

[Russell and Norvig, 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.

[SAE AS6983, 2019] SAE AS6983 (2019). *Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI*. SAE International.

[SAE J3016, 2018] SAE J3016 (2018). Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.

[Samoili et al., 2020] Samoili, S., Cobo, M. L., Gomez, E., De Prato, G., Martinez-Plumed, F., and Delipetrev, B. (2020). Ai watch. defining artificial intelligence. towards an operational definition and taxonomy of artificial intelligence. Joint Research Centre (Seville site).

[Sanders, 2009] Sanders, P. (2009). Algorithm engineering–an attempt at a definition. In *Efficient algorithms*, pages 321–340. Springer.

[Scharei et al., 2020] Scharei, K., Heidecker, F., and Bieshaar, M. (2020). Knowledge representations in technical systems–a taxonomy. *arXiv preprint arXiv:2001.04835*.

[Schölkopf et al., 2001] Schölkopf, B., Herbrich, R., Smola, A., Helmbold, D., and Williamson, B. (2001). Computational learning theory. *Lecture Notes in Computer Science*, 2111:416–426.

[Shhab et al., 2005] Shhab, A., Guo, G., and Neagu, D. (2005). A study on applications of machine learning techniques in data mining. *Department of Computing, University of Bradford, UK*.

[Statec, 2021] Statec (2021). Community survey on ict usage and e-commerce in enterprises. Technical report, Statec (Luxembourg).

[Stevenson, 2015] Stevenson, A., editor (2015). *Oxford Dictionary of English (3 ed.)*. Oxford University PressPrint.

[Sun, 2015] Sun, R. (2015). Artificial intelligence: Connectionist and symbolic approaches. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 35–40. Elsevier, Oxford, second edition edition.

[Szyperski et al., 2002] Szyperski, C., Gruntz, D., and Murer, S. (2002). *Component software: beyond object-oriented programming*. Pearson Education.

[Techopedia, 2018a] Techopedia (2018a). Recurrent neural network (rnn).

[Techopedia, 2018b] Techopedia (2018b). What is a generative adversarial network (gan)? - definition from techopedia.

[The Danish Government, 2019] The Danish Government (2019). National Strategy for Artificial Intelligence.

Titre: Taxonomie de l'IA de confiance - V2

Ce rapport propose une définition opérationnelle de l'intelligence artificielle de confiance à adopter dans le contexte du programme Confiance.ai, couvrant l'ensemble du cycle de vie d'un système critique basé sur l'IA. Cette définition opérationnelle est constituée d'une taxonomie concise et d'une liste de mots-clés qui caractérisent les domaines fondamentaux de l'intelligence artificielle de confiance, notamment les domaines suivants : ingénierie de l'IA, ingénierie des données, ingénierie des connaissances, ingénierie des algorithmes, ingénierie des logiciels et des systèmes, ingénierie de la sécurité, facteur humain et ingénierie cognitive. Ainsi, ce rapport propose une mise à jour de première version taxonomie de l'IA de confiance au regard du batch 2.

Title: Thrusworthy AI Taxonomy - V2

This report proposes an operational definition of trustworthy artificial intelligence to be adopted in the context of the Grand challenge "Confiance.ai", covering to the overall life cycle of an AI-based critical system. This operational definition is constituted by a concise taxonomy and a list of keywords that characterize the core domains of the trustworthy artificial intelligence including the following fields : AI Engineering, Data Engineering, Knowledge Engineering, Algorithm Engineering, Software and System Engineering, Safety Engineering, Human factor and Cognitive Engineering. Thus, this report proposes an update of the first version of the trustworthy AI taxonomy with respect to batch 2.

Our partners: