EU-U.S. Terminology and Taxonomy for Artificial Intelligence
First Edition

Introduction

The European Union (EU) and the United States (U.S.) are committed to cooperating on technologies and a digital transformation based on shared democratic values. The Trade and Technology Council (TTC) provides a platform for EU and U.S. policymakers and stakeholders to shape the future of transatlantic cooperation on Artificial Intelligence (AI).

As policy frameworks on AI emerge both in the EU and in the U.S., as well as in many other like-minded countries worldwide, the importance of aligning terminology and conceptual frameworks is becoming increasingly evident. Converging, interoperable approaches to defining and framing AI risks and trustworthiness are essential to enhance legal certainty, promote effective risk management, speed up the identification of emerging risks and reduce compliance costs and administrative burdens. This, in turn, is expected to foster innovation, maximising the benefits of AI systems and at the same time managing its risks. Ultimately the alignment of terminologies will help foster the EU-U.S. joint leadership in the development of an international standard for Trustworthy AI based on a mutual respect for human rights and democratic values.

As stated in the EU-U.S. Third Ministerial Statement, the first Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management (AI Roadmap) serves to inform the approaches to AI risk management and Trustworthy AI on both sides of the Atlantic, and advance collaborative approaches in international standards bodies related to AI. Following the Roadmap suggestions for concrete activities aimed at aligning EU and U.S. risk-based approaches, a group of experts engaged to prepare an initial draft AI terminologies and taxonomies. A total number of 65 terms were identified with reference to key documents from the EU and the U.S. (*see methodology below for more information*).

The identified terms reflect a shared technical, socio-technical and values-based understanding of AI systems between the EU and U.S. and will serve as a foundation for future definitions, as well as future transatlantic cooperation on AI terminology and taxonomy. This list should be considered as preliminary, to be further expanded and validated also with input from experts and stakeholders in the coming months.

Why AI Terminology Matters

AI terminology is pivotal to cooperation on AI in part due to the present momentum in the field, and due to the broader role of language in constructing and explaining scientific paradigms. Terminology is a necessary basis for technical standards and creates shared frames of reference between like-minded partners and across disciplines. Ultimately, different terminologies express distinct "technological cultures," thus revealing, through both alignment and divergence, the existence of gaps, unnecessary divergences and inconsistencies, and other points of departure for cooperation and collaboration.

The EU and U.S. understanding is based on the term "Trustworthy AI." According to the EU HLEG Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. According to the NIST AI Risk Management Framework (AI RMF), characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle.

In this context, different approaches to the development and governance of AI systems are currently competing at a regional and global level, resulting in distinct visions of technological systems based on the cultures of scientists and entrepreneurs as well as requirements and expectations from users, adopters, developers and lawmakers. The EU and

U.S. agree on the pursuit of a human-centric approach to AI: this requires that the terminology adopted to implement our shared approach to AI centres human, societal and environmental well-being, as well as the rule of law, human rights, democratic values and sustainable development.

Limitations and purpose of the terminology presented in this document

The list of terms presented in this document does not aim at achieving complete harmonisation or total alignment between the two legal systems. The EU and U.S. both recognise and respect their individual regulatory, social and cultural contexts, which in some instances may necessitate different definitions.

Furthermore, the list presented below does not include terms that are currently being discussed and defined in legislative processes in the EU and/or U.S., in order not to interfere with these.

Stakeholder Engagement

This document represents the first edition of the EU-U.S. Terminology and Taxonomy for Artificial Intelligence developed by the Working Group members according to the criteria and methodology presented below. This edition will be presented to AI experts and a broad community of stakeholders in the EU and the U.S. to receive feedback and contributions towards its enhancement and expansion. We therefore warmly encourage all stakeholders to share comments with the Working Group. Mechanisms for communication will be announced after the Fourth TTC Ministerial Meeting. These will be detailed separately.

Methodology

This list was built by the Working Group 1 experts from the EU and the U.S. in three steps. They initially defined a broad framework by agreeing on key criteria for selecting terms, largely based on existing official documents at the national and international level, as well as international standards documents and research publications. The selected terms were categorised into different clusters, and finally a list of terms are presented in this document. Below, these steps are described in more detail.

It must be noted that although many of the terms in this list can apply to several emerging technologies and technological systems, the terms in this list are only considered in the specific context of AI socio-technical systems.

1. Initial Step

a. The primary selection criteria were the following:

   i. Is this term essential to understanding a risk-based approach to AI?

   ii. Does the definition of this term serve to advance EU-U.S. cooperation on AI?

b. In defining terms, the experts turned to existing definitions found in widely-recognized documents such as academic literature, institutional references and the key EU-U.S. policy documents listed in the TTC Joint Roadmap for Trustworthy AI and Risk Management; and when needed tailored them to the context of AI.

2. Refined Step

Building upon the initial reference framework, the EU and U.S. experts further refined the selection of terms by undertaking the following exercises:

c. Jointly categorising terms as

   - **Foundational:** those terms which are essential to understanding the risk-based approach to the AI, and are relevant to and defined by both the EU and the U.S.
   - **Pending**: those terms whose definition is fixed or not changeable at this time due to legislative or other institutional processes occurring in either the EU or the U.S. These terms may be revisited in future revisions and efforts under the broader umbrella of the Joint AI Roadmap Implementation.

d. The EU and U.S. experts then compared and examined existing definitions and framing documents to find terms of greatest coherence or alignment between the EU and the U.S.

3. Proposed List of Terms

e. Through the process outlined above, the Working Group 1 experts have identified a preliminary list of terms which are believed to be essential to developing a transatlantic understanding of the risk-based approach to AI.

f. These terms reflect the shared understandings of AI systems between the EU and U.S. and may serve as a foundation for the ongoing work of the Working Group 1 and future transatlantic cooperation on AI terminology and taxonomy.

g. Annex A lists pending key terms that are currently involved in legislative or other institutional processes, and thus were excluded from the WG 1 efforts at this juncture.

List of Terms:

Note: *the references in this table are identified by a shorthand ID which is reflected in the references table in Annex B.*

1. Cluster: AI Lifecycle

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
adversarial machine learning (adversarial attack)	An input to a Machine Learning (ML) model that is purposely designed to cause a model to make a mistake in its predictions despite resembling a valid input to a human.	JRC	A practice concerned with the design of ML algorithms that can resist security challenges, the study of the capabilities of attackers, and the understanding of attack consequences.	Reznik,_Leon	A practice concerned with the design of ML algorithms that can resist security challenges, the study of the capabilities of attackers, and the understanding of attack consequences. Inputs in adversarial ML are purposely designed to make a mistake in its predictions despite resembling a valid input to a human.	Combination based on JRC and Reznik,_Leon
autonomy (autonomous AI system)		JRC	The system has a set of intelligence-based capabilities that allows it to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. Autonomous systems have a degree of self-government and self-directed behavior (with the human's proxy for decisions).	DOD_TEVV	Systems that maintain a set of intelligence-based capabilities to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. Autonomous systems have a degree of self-government and self-directed behaviour (with the human's proxy for decisions).	DOD_TEVV
big data	An all-encompassing term for any collection of data sets so large or complex that they are difficult to store, manage and process with conventional, non-scalable technology.	JRC	Extremely large data sets that are statistically analyzed to gain detailed insights. The data can involve billions of records and require substantial computer-processing power. Datasets are sometimes linked together to see how patterns in one domain affect other areas.	Brookings_Institution	An all-encompassing term for large, complex digital data sets that need equally complex technological means to be stored, analysed, managed and processed with substantial computing power. Datasets are sometimes linked together to see how patterns in one domain affect other areas. Data	Combination based on JRC and Brookings_Institution

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
			Data can be structured into fixed fields or unstructured as free-flowing information. The analysis of big datasets, often using AI, can reveal patterns, trends, or underlying relationships that were not previously apparent to researchers.		can be structured into fixed fields or unstructured as free-flowing information. The analysis of big datasets, often using AI, can reveal patterns, trends, or underlying relationships that were not previously apparent to researchers.	
classifier			A model that predicts (or assigns) class labels to data input.	own definition based on expertise	A model that predicts (or assigns) class labels to data input.	Own definition based on expertise.
data poisoning	Data poisoning occurs when an adversarial actor attacks an AI system training set, thus making the AI system learn something that it should not learn. Examples show that in some cases these data poisoning attacks on neural nets can be very effective, causing a significant drop in accuracy even with very little data poisoning. Other kinds of poisoning attacks do not aim to change the behaviour of the AI system, but rather they insert leverage to get the AI system to do what they want.	EU HLEG/ALTAI	Machine learning systems trained on user-provided data are susceptible to data poisoning attacks, whereby malicious users inject false training data with the aim of corrupting the learned model	Steinhardt,_Jacob	A type of security attack where malicious users inject false training data with the aim of corrupting the learned model, thus making the AI system learn something that it should not learn.	Combination based on HLEG/ALTAI and Steinhardt,_Jacob

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
deep learning	Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Deep learning architectures have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.	DL_1 DL_2	A subset of machine learning that relies on neural networks with many layers of neurons. In so doing, deep learning employs statistics to spot underlying trends or data patterns and applies that knowledge to other layers of analysis. Some have labeled this as a way to "learn by example" and a technique that "perform[s] classification tasks directly from images, text, or sound" and then applies that knowledge independently.Deep learning requires extensive computing power and labeled data, and is used in medical research, automated vehicles, electronics, and manufacturing, among other areas.	Brookings_Institution	A subset of machine learning based on artificial neural networks that employs statistics to spot underlying trends or data patterns and applies that knowledge to other layers of analysis. Some have labelled this as a way to "learn by example" and as a technique that "perform[s] classification tasks directly from images, text, or sound" and then applies that knowledge independently.	Combination based on DL_1, DL_2 and Brookings_Institution
differential privacy	Differential privacy is a meaningful and mathematically rigorous definition of privacy useful for quantifying and bounding privacy loss. Developed in the context of statistical disclosure control – providing accurate statistical information	Dwork_ECS	Differential privacy is a method for measuring how much information the output of a computation reveals about an individual. It is based on the randomised injection of "noise". Noise is a random alteration of data in a dataset so that values such as direct or indirect identifiers of individuals are harder to reveal. An important aspect of differential privacy is the	privacy-enhancing_technologies	Differential privacy is a method for measuring how much information the output of a computation reveals about an individual. It produces data analysis outcomes that are nearly equally likely, whether any individual is, or is not, included in the dataset. Its goal is to obscure the presence or absence of any individual (in a database), or small groups of individuals, while at the same	Combination based on privacy-enhancing_technologies and Dwork_ECS

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	about a set of respondents while protecting the privacy of each individual – the concept applies more generally to any private data set for which it is desirable to release coarse-grained information while keeping private the details. Informally, differential privacy requires the probability distribution on the published results of an analysis to be "essentially the same," independent of whether any individual opts in to or opts out of the data set. The probabilities are over the coin flips of the data analysis algorithm.		concept of "epsilon" or ε, which determines the level of added noise. Epsilon is also known as the "privacy budget" or "privacy parameter".		time preserving statistical utility.	
input data	Data provided to or directly acquired by an AI system on the basis of which the system produces an output.	EU AIA		IEEE_Soft_Vocab	Data provided to or directly acquired by an AI system on the basis of which the system produces an output.	Combination based on EU AIA and IEEE_Soft_Vocab
machine learning	Machine Learning (ML) is a branch of artificial intelligence (AI) and computer science which focuses on development	JRC	A general approach for determining models from data.	AI_Fairness_360	Machine Learning is a branch of artificial intelligence (AI) and computer science which focuses on development of systems	Combination based on JRC and AI_Fairness_360.

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	of systems that are able To learn and adapt without following explicit instructions imitating the way that humans learn, gradually improving its accuracy, by using algorithms and statistical models to analyse and draw inferences from patterns in data.				that are able to learn and adapt Without following explicit instructions imitating the way that humans learn, gradually improving its accuracy, by using algorithms and statistical models to analyse and draw inferences from patterns in data.	
model training	Process to establish or to improve the parameters of a machine learning model, based on a Machine Learning algorithm,by using training data.	ISO/IEC DIS 22989 Machine Learning	The phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.	C3.ai_Model_ Training	Process to establish or to improve the parameters of a machine learning model, based on a Machine Learning algorithm, by using training data.	ISO/IEC DIS22989 Machine Learning
model validation	Confirmation through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.	ISO/IEC DIS22989	The set of processes and activities intended to verify that models are performing as expected.	yields.io_mode l_validation	Confirmation through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.	ISO/IEC DIS22989
natural language processing	Information processing based upon natural language understanding and natural language generation. Discipline concerned with the way computers process natural language data.	ISO/IEC in JRC	A computer's attempt to “understand” spoken or written language. It must parse vocabulary, grammar, and intent, and allow for variation in language use. The process often involves machine learning.	Hutson,_Matth ew	The ability of a machine to process, analyse, and mimic human language, either spoken or written.	Own definition based on ISO/IEC in JRC and Hutson_Matthew

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
predictive analysis	Predictive analytics: this forward-looking technique aims to support the business in predicting what could happen by analysing backward-looking data. This involves the use of advanced data-mining and statistical techniques such as ML. The goal is to improve the accuracy of predicting a future event by analysing backward-looking data.	European Banking Authority	The organization of analyses of structured and unstructured data for inference and correlation that provides a useful predictive capability to new circumstances or data.	IEEE_Guide_IPA	The organisation of analyses of structured and unstructured data for inference and correlation that provides a useful predictive capability to new circumstances or data.	IEEE_Guide_IPA
profiling	'Profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.	GDPR	'Profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.	GDPR	'Profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.	GDPR
reinforcement learning	Machine Learning utilizing a reward function to optimize either a policy function or a value function by sequential interaction	ISO/IEC in JRC	A type of machine learning in which the algorithm learns by acting toward an abstract goal, such as "earn a high video game score" or "manage a factory	Hutson,_Matthew	A type of machine learning in which the algorithm learns by acting toward an abstract goal, such as "earn a high video game score" or "manage a factory efficiently." During	Combination based on Hutson,_Matthew and ISO/IEC in JRC

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	with an environment. Note 1 to entry: Policy functions and value functions express a strategy that is learned by the environment. Note 2 to entry: The environment can be any stateful model.		efficiently." During training, each effort is evaluated based on its contribution toward the goal.		training, each effort is evaluated based on its contribution toward the goal.	
structured data			Data that has a predefined data model or is organized in a predefined way.	NIST_1500	Data that has a predefined data model or is organised in a predefined way.	NIST_1500
unstructured data			Data that does not have a predefined data model or is not organized in a predefined way.	Own definition based on NIST_1500	Data that does not have a predefined data model or is not organised in a predefined way.	Own definition based on NIST_1500
synthetic data	Synthetic data is artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data. This means that synthetic data and original data should deliver very similar results when undergoing the same statistical analysis. The degree to which synthetic data is an accurate proxy for the original data is a measure of the utility of the method and the model. The generation process, also called synthesis, can be performed using	EDPS_SD	Synthetic data can mean many different things depending upon the way they are used. Sometimes, as in computer programming, the term means data that are completely simulated for testing purposes. Other times, as in statistics, the term means combining data, often from multiple sources, to produce estimates for more granular populations than any one source can support. An example of this usage is the U.S. Census Bureau's Small Area Income and Poverty Estimates. In data confidentiality applications, synthetic data are modeled statistical outputs released in a format that closely resembles the confidential	U.S. Census	Synthetic data is generated from data/processes and a model that is trained to reproduce the characteristics and structure of the original data aiming for similar distribution. The degree to which synthetic data is an accurate proxy for the original data is a measure of the utility of the method and the model.	Own definition based on EDPS_SD

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	different techniques, such as decision trees, or deep learning algorithms. Synthetic data can be classified with respect to the type of the original data: the first type employs real datasets, the second employs knowledge gathered by the analysts instead, and the third type is a combination of these two. Generative Adversarial Networks (GANs) were introduced recently and are commonly used in the field of image recognition. They are generally composed of two neural networks training each other iteratively. The generator network produces synthetic images that the discriminator network tries to identify as such in comparison to real images.		data format. Synthetic data can be disaggregated to the individual- or business-record level, or aggregated into tabular format.			
transfer learning			A technique in machine learning in which an algorithm learns to perform one task, such as recognizing cars, and builds on that knowledge when learning a	Hutson,_Matthew	A technique in machine learning in which an algorithm learns to perform one task, such as recognizing cars, and builds on that knowledge when learning a different but related task, such as recognizing cats.	Hutson,_Matthew

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
			different but related task, such as recognizing cats.			
supervised learning	Machine learning that makes use of labelled data during training	ISO/IEC DIS22989			Machine learning that makes use of labelled data during training.	ISO/IEC DIS22989
unsupervised learning	Machine learning that makes use of unlabelled data during training.	ISO/IEC in JRC	Algorithms, which take a set of data consisting only of inputs and then they attempt to cluster the data objects based on the similarities or dissimilarities in them.	Reznik,_Leon	Machine learning that makes use of unlabelled data during training.	ISO/IEC in JRC

2. Cluster: Measurement

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
(AI) accuracy	The goal of an AI model is to learn patterns that generalize well for unseen data. It is important to check if a trained AI model is performing well on unseen examples that have not been used for training the model. To do this, the model is used to predict the answer on the test dataset and then the predicted target is compared to the actual answer. The concept of accuracy is used to evaluate the predictive capability of the AI model. Informally, accuracy is the fraction of predictions the model got right. A number of metrics are used in machine learning (ML) to measure the predictive accuracy of a model. The choice of the accuracy metric to be used depends on the ML task.	EU HLEG/ALTAI	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.	OECD	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. The goal of an AI model is to learn patterns that generalise well for unseen data. It is important to check if a trained AI model is performing well on unseen examples that have not been used for training the model. To do this, the model is used to predict the answer on the test dataset and then the predicted target is compared to the actual answer. The concept of accuracy is used to evaluate the predictive capability of the AI model. Informally, accuracy is the fraction of predictions the model got right. A number of metrics are used in machine learning (ML) to measure the predictive accuracy of a model. The choice of the accuracy metric to be used depends on the ML task.	Combination based on EU HLEG/ALTAI and OECD.
Test			Technical operation to determine one or more characteristics of or to evaluate the performance of a given product, material, equipment, organism,	NSCAI	Technical operation to determine one or more characteristics of or to evaluate the performance of a given product, material, equipment, organism, physical	NSCAI

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
			physical phenomenon, process or service according to a specified *procedure*. OR Activity in which a system or component is executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component		phenomenon, process or service according to a specified procedure.	
Evaluation			Systematic determination of the extent to which an entity meets its specified criteria	NSCAI	Systematic determination of the extent to which an entity meets its specified criteria.	NSCAI
Verification			Provides evidence that the system or system element performs its intended functions and meets all performance requirements listed in the system performance specification	NSCAI	Provides evidence that the system or system element performs its intended functions and meets all performance requirements listed in the system performance specification.	NSCAI
Validation			Confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled	NSCAI	Confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled.	NSCAI
Test and Evaluation, Verification and Validation (TEVV)			A framework for assessing, incorporating methods and metrics to determine that a technology or system satisfactorily meets its design specifications and requirements, and that it is sufficient for its intended use.	NSCAI	A framework for assessing, incorporating methods and metrics to determine that a technology or system satisfactorily meets its design specifications and requirements, and that it is sufficient for its intended use.	NSCAI

3. Cluster: Technical System Attributes

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
adaptive learning	An adaptive AI is a system that changes its behaviour while in use. Adaptation may entail a change in the weights of the model or a change in the internal structure of the model itself. The new behaviour of the adapted system may produce different results than the previous system for the same inputs.	ISO/IEC DIS 22989 Trustworthiness	Updating predictive models online during their operation to react to concept drifts	Gama_Joao	An adaptive AI is a system that changes its behaviour while in use. Adaptation may entail a change in the weights of the model or a change in the internal structure of the model itself. The new behaviour of the adapted system may produce different results than the previous system for the same inputs.	ISO/IEC DIS 22989 Trustworthiness
algorithm	An algorithm consists of a set of instructions or steps used to solve a problem (e.g., it does not include the data). The algorithm can be abstract and implemented in different programming languages and software libraries.	JRC	A set of step-by-step instructions. Computer algorithms can be simple (if it's 3 p.m., send a reminder) or complex (identify pedestrians).	Huston,_Matthew	An algorithm consists of a set of step-by-step instructions to solve a problem (e.g., not including data). The algorithm can be abstract and implemented in different programming languages and software libraries.	Combination based on JRC and Huston_Matthew
classification			When the output is one of a finite set of values (such as sunny, cloudy or rainy), the learning problem is called classification, and is called Boolean or binary classification if there are only two values.	AIMA	A classification system is a set of "boxes" into which things are sorted. Classifications are consistent, have unique classificatory principles, and are mutually exclusive. In AI design, when the output is one of a finite set of values (such as sunny, cloudy or rainy), the learning problem is called classification, and is called Boolean or binary	Own definition based on AIMA and definition of classification by Bowker and Star.

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
					classification if there are only two values.	
federated learning	Federated learning is a relatively new way of developing machine-learning models where each federated device shares its local model parameters instead of sharing the whole dataset used to train it. The federated learning topology defines the way parameters are shared. In a centralised topology, the parties send their model parameters to a central server that uses them to train a central model which in turn sends back updated parameters to the parties. In other topologies, such as the peer-to-peer or hierarchical one, the parties share their parameters with a subset of their peers. Federated learning is a potential solution for developing machine-learning models that require huge or very disperse datasets. However, it is not a one-size-fits-all machine learning scenarios	EDPS_FL	a learning model which addresses the problem of data governance and privacy by training algorithms collaboratively without transferring the data to another location.	Public_Health_and_Informatics_MIE_2021	Federated learning is a machine learning model which addresses the problem of data governance and privacy by training algorithms collaboratively without transferring the data to another location. Each federated device shares its local model parameters instead of sharing the whole dataset used to train it and the federated learning topology defines the way parameters are shared.	Own definition based on combination of EDPS_FL and Public_Health_and_Informatics_MIE_2021

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
generative adversarial network (GAN)			Generative Adversarial Networks, or GANs for short, are an approach to generative modeling using deep learning methods, such as convolutional neural networks. Generative modeling is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.	Brownlee,_Jason_GAN	Generative Adversarial Networks, or GANs for short, are an approach to generative modelling using deep learning methods, such as convolutional neural networks. Generative modelling is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.	Brownlee,_Jason_GAN

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
human values for AI	"Respect for rule of law, human rights and democratic values" The European Union declares the fundamental EU values to be the ones "common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail". They are: human dignity, freedom, democracy, equality, rule of law, and human rights.	EP Human Rights Fact Sheet	Artificial intelligence systems use data we generate in our daily lives and as such are a mirror of our interests, weaknesses, and differences. Artificial intelligence, like any other technology, is not value-neutral. Understanding the values behind the technology and deciding on how we want our values to be incorporated in AI systems requires that we are also able to decide on how and what we want AI to mean in our societies. It implies deciding on ethical guidelines, governance policies, incentives, and regulations. And it also implies that we are aware of differences in interests and aims behind AI systems developed by others according to other cultures and principles. **See note.*	Virginia_Dignum_Responsibility_and_Artificial_Intelligence	Values are idealised qualities or conditions in the world that people find good. AI systems are not value-neutral. The incorporation of human values into AI systems requires that we identify whether, how and what we want AI to mean in our societies. It implies deciding on ethical principles, governance policies, incentives, and regulations. And it also implies that we are aware of differences in interests and aims behind AI systems developed by others according to other cultures and principles. The EU and U.S. are committed to the development of Trustworthy AI systems based on shared democratic values including the respect for the rule of law and human rights.	Own definition based on EU and U.S. values and Brey
human-centric AI	AI is not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being.	HLEG AI, Ethics Guidelines for Trustworthy AI.			An approach to AI that prioritises human ethical responsibility, dynamic qualities, understanding and meaning. It encourages the	Own definition based on Hasselbalch, G. (2021) Data Ethics of Power - A Human

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoy a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come.				empowerment of humans in design, use and implementation of AI systems. Human-Centric AI systems are built on the recognition of a meaningful human-technology interaction. They are designed as components of socio-technical environments in which humans assume meaningful agency. Human-Centric AI is not designed as an end in itself, but as tools to serve people with the ultimate aim of increasing human and environmental well-being with respect for the rule of law, human rights, democratic values and sustainable development.	Approach to Big Data and AI, Edward Elgar; HLEG Ethics Guidelines for Trustworthy AI.
language model			A language model is an approximative description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments.	Gustavii,_Ebba	A language model is an approximative description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments.	Gustavii,_Ebba

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
large language model (LLM)			A class of language models that use deep-learning algorithms and are trained on extremely large textual datasets that can be multiple terabytes in size. LLMs can be classed into two types: generative or discriminatory. Generative LLMs are models that output text, such as the answer to a question or even writing an essay on a specific topic. They are typically unsupervised or semi-supervised learning models that predict what the response is for a given task. Discriminatory LLMs are supervised learning models that usually focus on classifying text, such as determining whether a text was made by a human or AI.	AI_Assurance_2022	A class of language models that use deep-learning algorithms and are trained on extremely large textual datasets that can be multiple terabytes in size. LLMs can be classed into two types: generative or discriminatory. Generative LLMs are models that output text, such as the answer to a question or even writing an essay on a specific topic. They are typically unsupervised or semi-supervised learning models that predict what the response is for a given task. Discriminatory LLMs are supervised learning models that usually focus on classifying text, such as determining whether a text was made by a human or AI	AI_Assurance_2022
model	The workflow of an AI model shows the phases needed to build the model and their interdependencies. Typical phases are: Data collection and preparation, Model development, Model training, Model accuracy evaluation, Hyperparameters' tuning, Model usage, Model maintenance, Model versioning. These stages are usually iterative: one	EU HLEG/ALTAI	A function that takes features as input and predicts labels as output.	AI_Fairness_360	A function that takes features as input and predicts labels as output. Typical phases of an AI model's work flow are: Data collection and preparation, Model development, Model training, Model accuracy evaluation, Hyperparameters' tuning, Model usage, Model maintenance, Model versioning.	Combination based on EU HLEG/ALTAI and AI_Fairness_360

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	may need to reevaluate and go back to a previous step at any point in the process.					
neural network	Network of two or more layers of neurons connected by weighted links with adjustable weights, which takes input data and produces an output. Note 1 to entry: Whereas some neural networks are intended to simulate the functioning of biological neurons in the nervous system, most neural networks are used in artificial intelligence as realizations of the connectionist model.	ISO/IEC in JRC	Also known as artificial neural network, neural net, deep neural net; a computer system inspired by living brains.	Ranschaert,_Erik	A computer system inspired by living brains, also known as artificial neural network, neural net, or deep neural net. It consists of two or more layers of neurons connected by weighted links with adjustable weights, which takes input data and produces an output. Whereas some neural networks are intended to simulate the functioning of biological neurons in the nervous system, most neural networks are used in artificial intelligence as realisations of the connectionist model.	Own definition based on ISO/IEC in JRC and Ranschaert,_Erik
scalability			The ability to increase or decrease the computational resources required to execute a varying volume of tasks, processes, or services.	IEEE_Guide_IP A	The ability to increase or decrease the computational resources required to execute a varying volume of tasks, processes, or services.	IEEE_Guide_IPA
socio-technical system	Technology is always part of society, just like society is always part of technology. This also means that one cannot understand one without the other. Technology is not only design and material appearance but also sociotechnical; that is, a complex process	(Hasselbach (2021) based on Hughes, 1987, 1983; Bijker et al., 1987; Misa, 1988, 1992, 2009; Bijker & Law, 1992; Edwards, 2002;	how humans interact with technology within the broader societal context.		Technology is always part of society, just like society is always part of technology. This also means that one cannot understand one without the other. Technology is not only design and material appearance but also sociotechnical; that is, a complex process constituted by diverse social, political, economic, cultural and technological factors.	Hasselbalch (2021), based on Hughes, 1987, 1983; Bijker et al., 1987; Misa, 1988, 1992, 2009; Bijker & Law, 1992; Edwards, 2002; Harvey et al., 2017 etc.

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	constituted by diverse social, political, economic, cultural and technological factors.	Harvey et al., 2017 etc.).				
technical interoperability			The ability of software or hardware systems or components to operate together successfully with minimal effort by end user.	SP1011	The ability of software or hardware systems or components to operate together successfully with minimal effort by an end user.	SP1011
value sensitive design (values-by-design or ethics-by-design)			A theoretically grounded approach to the design of technology that accounts for human values in a principled and systematic manner throughout the design process.	Friedman_et_al_2017	A theoretically grounded approach to the design of technology that accounts for human values in a principled and systematic manner throughout the design process.	Friedman_et_al_2017

4. Cluster: Governance

Term	EU definition	Source	U.S. definition	Source	Final definition	Final Source
auditability of an AI system	Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and Intellectual Property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enable the system's auditability.	EU HLEG/ALTAI	Systematic, independent, documented process for obtaining records, statements of fact, or other relevant information and assessing them objectively, to determine the extent to which specified requirements are fulfilled.	IEEE_Soft_Vocab	Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and Intellectual Property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enable the system's auditability.	EU HLEG/ALTAI
standard	"Standards are a set of institutionalised agreed upon-rules for the production of (textual or material) objects. They are released by international organizations and ensure quality and safety and set product or services' specifications. Standards are the result of negotiations among various stakeholders and are institutionalised and thus difficult to change."	loosely based Bowker and Star			Standards are a set of institutionalised agreed upon-rules for the production of (textual or material) objects. They are released by international organisations and ensure quality and safety and set product or services' specifications. Standards are the result of negotiations among various stakeholders and are institutionalised and thus difficult to change.	Loosely based on Bowker and Star

5. Cluster: Trustworthy

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
accessibility	Extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (which includes direct use or use supported by assistive technologies).	EU HLEG/ALTAI			Extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (which includes direct use or use supported by assistive technologies).	EU HLEG/ALTAI
accountability	This term refers to the idea that one is responsible for their action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons. Accountability has several dimensions. Accountability is sometimes required by law. For example, the General Data Protection Regulation (GDPR) requires organisations that process personal data to ensure security measures are in place to prevent data breaches and report if these fail. But accountability might also express an ethical standard, and fall short of	EU HLEG/ALTAI	Accountability relates to an allocated responsibility. The responsibility can be based on regulation or agreement or through assignment as part of delegation; 2) For systems, a property that ensures that actions of an entity can be traced uniquely to the entity; 3) In a governance context, the obligation of an individual or organization to account for its activities, for completion of a deliverable or task, accept the responsibility for those activities, deliverables or tasks, and to disclose the results in a transparent manner.	ISO/IEC_TS_5723:2022	Accountability relates to an allocated responsibility. The responsibility can be based on regulation or agreement or through assignment as part of delegation. In a systems context, accountability refers to systems and/or actions that can be traced uniquely to a given entity. In a governance context, accountability refers to the obligation of an individual or organisation to account for its activities, to complete a deliverable or task, to accept the responsibility for those activities, deliverables or tasks, and to disclose the results in a transparent manner.	Combination based on EU HLEG/ALTAI and ISO/IEC_TS_5723:2022

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	legal consequences. Some tech firms that do not invest in facial recognition technology in spite of the absence of a ban or technological moratorium might do so out of ethical accountability considerations.					
AI (or algorithmic) bias	AI (or algorithmic) bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others. Bias can emerge due to many factors, including but not limited to the design of the algorithm or the unintended or unanticipated use or decisions relating to the way data is coded, collected, selected or used to train the algorithm. Bias can enter into algorithmic systems as a result of pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; or by being used in unanticipated contexts or by audiences who are not considered in the	EU HLEG/ALTAI	A systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage.	AI_Fairness_360	Harmful AI bias describes systematic and repeatable errors in AI systems that create unfair outcomes, such as placing. privileged groups at systematic advantage and unprivileged groups at systematic disadvantage. Different types of bias can emerge and interact due to many factors, including but not limited to, human or system decisions and processes across the AI lifecycle. Bias can be present in AI systems resulting from pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; by being used in unanticipated contexts; or by non-representative design specifications.	Combination based on EU HLEG/ALTAI and AI_Fairness_360

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	software's initial design. AI bias is found across platforms, including but not limited to search engine results and social media platforms, and can have impacts ranging from inadvertent privacy violations to reinforcing social biases of race, gender, sexuality, and ethnicity.					
attack	Model inversion refers to a kind of attack to AI models, in which the access to a model is abused to infer information about the training data. So, model inversion turns the usual path from training data into a machine-learned model from a one-way one to a two-way one, permitting the training data to be estimated from the model with varying degrees of accuracy. Such attacks raise serious concerns given that training data usually contain privacy-sensitive information.	EU HLEG/ALTAI	Action targeting a learning system to cause malfunction.	NISTIR_8269_Draft	Action targeting a learning system to cause malfunction.	NISTIR_8269_Draft
chatbot (conversational bot)	A computer program designed to simulate conversation with a human user, usually over the internet; especially one used	Oxford English Dictionary	Conversational agent that dialogues with its user (for example: empathic robots available to patients, or automated conversation services in customer relations).	COE_AI_Glossary	A computer program designed to simulate conversation with a human user, usually over the internet; especially one used	Oxford English Dictionary

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	to provide information or assistance to the user as part of an automated service.				to provide information or assistance to the user as part of an automated service.	
discrimination	Pre-existing bias comes from the outside of the computer system. It can be individual or social, and it already exists in social contexts and in the personal biases and attitudes held by the developers of the system. This type of bias is embedded in a computer system either explicitly and deliberately or implicitly and undeliberately by institutions or individuals. Technical bias comes from technical constraints or limitations, like imperfections in pseudorandom number generation that, for example, systematically favour those at the end of a database. Finally, emergent bias appears in the context of use of a computer system." The aim of non-discrimination law is to	EU LEX	Disadvantageous treatment of a person based on belonging to a category rather than on individual merit.	Žliobaitė_Indrė	Unequal treatment of a person based on belonging to a category rather than on individual merit. Discrimination can be a result of societal, institutional and implicitly held individual biases or attitudes that get captured in processes across the AI lifecycle, including by AI actors and organisations, or represented in the data underlying AI systems. Discrimination biases can also emerge due to technical limitations in hardware or software, or the use of an AI system that, due to its context of application, does not treat all groups equally. Discriminatory biases can also emerge in the very context in which the AI system is used. As many forms of biases are systemic and implicit, they are not easily controlled or mitigated and require specific governance and other similar approaches"	Own definition loosely based on Friedman, B. and Nissenbaum, H.; and Schwartz, R. et al.

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	allow all individuals an equal and fair chance to access opportunities available in a society. This means that individuals or groups of individuals which are in comparable situations should not be treated less favourably simply because of a particular characteristic such as their sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation.					
evasion	Evasion is one of the most common attacks on machine learning models (ML) performed during production. It refers to designing an input, which seems normal for a human but is wrongly classified by ML models. A typical example is to change some pixels in a picture before uploading, so that the image recognition system fails to classify the result.	EU HLEG/AL TAI	In Evasion Attacks, the adversary solves a constrained optimization problem to find a small input perturbation that causes a large change in the loss function and results in output misclassification.	tabassi_adversarial_2019	In Evasion Attacks, the adversary solves a constrained optimization problem to find a small input perturbation that causes a large change in the loss function and results in output misclassification.	tabassi_adversarial_2019
fault tolerance	Fault tolerance is the property that enables a system to continue operating properly in the event of the failure of (or one or more faults within) some of its components.	EU HLEG/AL TAI	The ability of a system or component to continue normal operation despite the presence of hardware or software faults	SP1011	The ability of a system or component to continue normal operation despite the presence of hardware or software faults.	SP1011

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	If its operating quality decreases at all, the decrease is proportional to the severity of the failure, as compared to a naively designed system, in which even a small failure can cause total breakdown. Fault tolerance is particularly sought after in high-availability or safety- critical systems. Redundancy or duplication is the provision of additional functional capabilities that would be unnecessary in a fault-free environment. This can consist of backup components that automatically 'kick in' if one component fails.					
feedback loop			describes the process of leveraging the output of an AI system and corresponding end-user actions in order to retrain and improve models over time. The AI-generated output (predictions or recommendations) are compared against the final decision (for example, to perform work or not) and provides feedback to the model, allowing it to learn from its mistakes.	C3.ai_feedback_loop	Feedback loop describes the process of leveraging the output of an AI system and corresponding end-user actions in order to retrain and improve models over time. The AI-generated output (predictions or recommendations) are compared against the final decision (for example, to perform work or not) and provides feedback to the model, allowing it to learn based on its results.	Own definition based on C3.ai_feedback_loop

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
harmful bias			Harmful bias can be either conscious or unconscious. Unconscious, also known as implicit bias, involves associations outside conscious awareness that lead to a negative evaluation of a person on the basis of characteristics such as race, gender, sexual orientation, or physical ability.3,14 Discrimination is behavior; discriminatory actions perpetrated by individuals or institutions refer to inequitable treatment of members of certain social groups that results in social advantages or disadvantages	humphrey_addressing_2020	Harmful bias can be either conscious or unconscious. Unconscious, also known as implicit bias, involves associations outside conscious awareness that lead to a negative evaluation of a person on the basis of characteristics such as race, gender, sexual orientation, or physical ability. Discrimination is behaviour; discriminatory actions perpetrated by individuals or institutions refer to inequitable treatment of members of certain social groups that results in social advantages or disadvantages. AI systems can reinforce harmful bias when trained on prejudiced or unrepresentative data. Most often harmful bias is unintended by developers and adopters of AI. AI actors can design AI systems to mitigate harmful bias.	humphrey_addressing_2020
human rights impact assessment	The rights people are entitled to simply because they are human beings, irrespective of their citizenship, nationality, race, ethnicity, language, gender, sexuality, or abilities; human rights become enforceable when they are codified as conventions, covenants, or treaties.	DIHR	Impact assessment definition - a risk management tool that seeks to ensure an organization has sufficiently considered a system's relative benefits and costs before implementation. In the context of AI, an impact assessment helps to answer a simple question: alongside this system's intended use, for whom could it fail?		An human rights impact assessment (HRIA) of AI identifies, understands and assesses the impact of the AI system on human rights, such as but not limited to, the right to privacy or non-discrimination. AI systems can pose risks to, as well as enhance, individual human rights.	Own definition based on DIHR

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
opacity	The opacity refers to the lack of transparency on the process by which AI system reaches a result. An AI system can be transparent (or conversely opaque) in three different ways: with respect to how exactly the AI system functions as a whole (functional transparency); how the algorithm was realized in code (structural transparency) and how the program actually run in a particular case, including the hardware and input data (run transparency). Algorithms often no longer take the form of more or less easily readable code, but instead resemble a 'black-box'. This means that while it maybe be possible to test the algorithm as to its effects, but not to understand how those effects have been achieved. Some AI systems lack transparency because the rules followed, which lead from input to output, are not fully prescribed by a human. Rather, is some cases, the algorithm is set to learn from data in order	EU AIA (Impact Assessment of the AI Act, Annex 5.2)	[to receive] the output of [an] algorithm (the classification decision) [and to not] have any concrete sense of how or why a particular classification has been arrived at from inputs. When AI system processes, functions, output or behavior are unavailable or incomprehensible to all stakeholders - usually an antonym for transparency.	Jenna_Burrell	When AI system processes, functions, output or behaviour are unavailable or incomprehensible to all stakeholders – usually an antonym for transparency.	Jenna_Burrell

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	to arrive at a pre-defined output in the most efficient way, which might not be representable by rules which a human could understand. As a result, AI systems are often opaque in a way other digital systems are not ('the so called black box effect'). Independently from technical characteristics, a lack of transparency can also stem from systems relying on rules and functionalities that are not publicly accessible and of which a meaningful and accurate description is not publicly accessible. The complexity and lack of transparency (opacity of AI) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights.					
red-team	Red teaming is the practice whereby a red team or independent group challenges an organisation to improve its effectiveness by assuming an adversarial	EU HLEG/AL TAI	A group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture. The Red Team's objective is to improve enterprise	CSRC	A group of people authorised and organised to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture. It is often used to help identify	Combination based on EU HLEG/ALTAI and CSRC

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	role or point of view. It is often used to help identify and address potential security vulnerabilities.		cybersecurity by demonstrating the impacts of successful attacks and by demonstrating what works for the defenders (i.e., the Blue Team) in an operational environment. Also known as Cyber Red Team.		and address potential security vulnerabilities.	
reliability	An AI system is said to be reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested earlier.	EU HLEG/ALTAI	Reliability refers to the closeness of the initial estimated value(s) to the subsequent estimated values.	OECD	An AI system is said to be reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested earlier.	EU HLEG/ALTAI
resilience			The ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or incidents. The ability of a system to adapt to and recover from adverse conditions.	NISTIR_8269_Draft	The ability of an AI system to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or incidents. The ability of a system to adapt to and recover from adverse conditions.	NISTIR_8269_Draft
robustness (robust AI)	Robustness of an AI system encompasses both its technical robustness (appropriate in a given context, such as the application domain or life cycle phase) and as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is	EU HLEG/ALTAI	ability of a system to maintain its level of performance under a variety of circumstances	ISO/IEC_TS_5723:2022	Robustness of an AI system encompasses both its technical robustness (ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur.	Own definition based EU HLEG/ALTAI and ISO/IEC TS_5723:2022

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	crucial to ensure that, even with good intentions, no unintentional harm can occur. Robustness is the third of the three components necessary for achieving Trustworthy AI.					
safety	AI safety is described as mitigating accident risks from machine learning systems. "The problem of accidents in machine learning systems. We define accidents as unintended and harmful behaviour that may emerge from machine learning systems when we specify the wrong objective function, are not careful about the learning process, or commit other machine learning related implementation errors."	Amodei	AI systems should not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.	ISO ISO/IEC TS 5723:2022	AI systems should not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.	ISO ISO/IEC TS 5723:2022
security			The protection mechanisms, design and maintenance of an AI system and infrastructure's AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms.	NIST_AI_RMF_1.0	The protection mechanisms, design and maintenance of an AI system and infrastructure's AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms.	NIST_AI_RMF_1.0
systemic bias			Systemic biases result from procedures and practices of particular institutions that operate in ways which result in certain social groups being	D. Chandler and R. Munday	Systemic bias is a social consistent structure of harmful bias that is systemically reinforced in institutions, cultural perception and socio-	Own definition loosely based on D. Chandler and R. Munday

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
			advantaged or favored and others being disadvantaged or devalued. This need not be the result of any conscious prejudice or discrimination but rather of the majority following existing rules or norms.		technical infrastructures. AI systems can reinforce systemic biases by reproducing the discriminatory effects of systemic biases when deployed in socially important institutions, cultural production or in societal infrastructures.	
traceability	Ability to track the journey of a data input through all stages of sampling, labelling, processing and decision making.	EU HLEG/ALTAI	Ability to trace the history, application or location of an entity by means of recorded identification. ["Chain of custody" is a related term.] Alternatively, traceability is a property of the result of a measurement or the value of a standard whereby it can be related with a stated uncertainty, to stated references, usually national or international standards, i.e. through an unbroken chain of comparisons. In this context, the standards referred to here are measurement standards rather than written standards.	UNODC_Glossary_QA_GLP	Ability to track the journey of a data input through all stages of sampling, labelling, processing and decision making.	EU HLEG/ALTAI
Trustworthy AI	Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since,	EU HLEG/ALTAI	Characteristics of Trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.	NIST_AI_RMF_1.0	Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.	Own definition based on combination of EU HLEG/ALTAI and NIST_AI_RMF_1.0

Term	EU definition	Reference	U.S. definition	Reference	Final definition	Final Source
	even with good intentions, AI systems can cause unintentional harm. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle.				Characteristics of Trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle.	

Annex A. Pending Terms:

1. Artificial general intelligence (AGI)
2. Artificial intelligence (AI)
3. Biometric categorisation system
4. Biometric data
5. Conformity
6. Data governance
7. Data quality
8. Deepfake
9. Deployment
10. Developer
11. Distributor
12. Documentation
13. Emotion recognition system
14. Foundation models
15. Fundamental rights impact assessment
16. Generative artificial intelligence
17. Harm
18. Importer
19. Incident
20. Instructions for use (usability)
21. Life cycle of an AI system
22. Minimisation
23. Misuse
24. Model risk management
25. Operator
26. Provider
27. Pseudo-anonymization (pseudonymization)
28. Recall of an AI system
29. Rectification
30. Remote biometric identification system
31. Risk
32. Risk control
33. Sandbox
34. Sensitive data
35. Subliminal techniques

36. Substantial modification
37. Testing data
38. Testing in real world conditions
39. Training data
40. User
41. Validation

Annex B. References

EU References

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
Amodei	Computational Disciplines	Amodei et al.	IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems				2016	https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/eadv2_glossary.pdf	
Bowker and Star	Sorting Things out: Classification and Its Consequences.	Bowker, G.C. and Star S.L.	Inside Technology, MIT.				2000		
Brey	Values in technology and disclosive ethics	Brey, P.	L. Floridi (ed.) The Cambridge Handbook of Information and Computer Ethics, Cambridge University Press			41–58	2010		
DIHR	Introduction to human rights impact assessment		The Danish Institute for Human Rights					https://www.humanrights.dk/tools/human-rights-impact-assessment-guidance-toolbox/introduction-human-	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
								rights-impact-assessment	
DL_1	Representation Learning: A Review and New Perspectives	Bengio, Y; Courville, A; Vincent, P.	IEEE Transactions on Pattern Analysis and Machine Intelligence				2013	https://arxiv.org/pdf/1206.5538.pdf	
DL_2	Deep Learning in Neural Networks: An Overview	Schmidhuber, J.	Neural Networks				2015	https://arxiv.org/abs/1404.7828	
Dwork_ECS	Differential Privacy	Dwork, C.	Encyclopedia of Cryptography and Security				2011	https://doi.org/10.1007/978-1-4419-5906-5_752	
EDPS_FL	Federated Learning	Lareo, Xabier	European Data Protection Supervisor					https://edps.europa.eu/press-publications/publications/techsonar/federated-learning_en	
EDPS_SD	What are Synthetic Data?	Riemann, Robert	European Data Protection Supervisor					https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
EP Human Rights Fact Sheet	Human Rights	European Parliament	European Parliament				2022	https://www.europarl.europa.eu/factsheets/en/sheet/165/human-rights	
EU AIA	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						2021	https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206	
EU HLEG/ALTAI	Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment	European Union High Level Expert Group on Artificial Intelligence					2020	https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment	
EU LEX	Non-discrimination (the principle of)	European Union Treaty on the Functioning of the EU (TFEU)						https://eur-lex.europa.eu/EN/legal-content/glossary/non-discrimination-	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
								the-principle-of.html	
European Banking Authority	EBA Report on Big Data and Advanced Analytics	European Banking Authority					2020	https://www.eba.europa.eu/sites/default/documents/files/document_library//Final%20Report%20on%20Big%20Data%20and%20Advanced%20Analytics.pdf?retry=1	
Friedman, B. and Nissenbaum, H.	Bias in Computer Systems	Friedman, B., Nissenbaum, H	ACM Transactions on Information Systems	14	3	330–47	1996		
GDPR	Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free							https://eur-lex.europa.eu/eli/reg/2016/679/oj	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
	movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)								
Hasselbach		Hasselbach, G.	Data Ethics of Power: A Human Approach in the Big Data and AI Era				2021	https://www.e-elgar.com/shop/gbp/data-ethics-of-power-9781802203103.html	Based on: Hughes, 1987, 1983; Bijker et al., 1987; Misa, 1988, 1992, 2009; Bijker & Law, 1992; Edwards, 2002; Harvey et al., 2017 etc.
HLEG AI, Ethics Guidelines for Trustworthy AI	Ethics guidelines for trustworthy AI						2019	https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai	
ISO/IEC DIS 22989 Machine Learning	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology		ISO					https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:22989:dis:ed-1:v1:en	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
ISO/IEC DIS 22989 Trustworthiness	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology		ISO				2022	https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:22989:dis:ed-1:v1:en	
ISO/IEC DIS22989	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology						2022	https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:22989:dis:ed-1:v1:en	
ISO/IEC in JRC	Glossary of human-centric artificial intelligence	Estevez Almenzar Marina; Fernandez Llorca David; Gomez Gutierrez Emilia; Martinez Plumed Fernando	EU Joint Research Centre				2022	https://publications.jrc.ec.europa.eu/repository/handle/JRC129614	
JRC	Glossary of human-centric artificial intelligence	Estevez Almenzar Marina; Fernandez Llorca David; Gomez Gutierrez	EU Joint Research Centre				2022	https://publications.jrc.ec.europa.eu/repository/handle/JRC129614	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
		Emilia; Martinez Plumed Fernando							
Oxford English Dictionary	Chatbot	Oxford English Dictionary	Oxford English Dictionary				2020	https://www.oed.com/view/Entry/88357851?redirectedFrom=chatbot#eid	
Schneiderman	Human-Centered AI	Schneiderman, B.	Oxford University Press				2022	https://global.oup.com/academic/product/human-centered-ai-9780192845290?cc=fr&lang=en&	

U.S. References

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
AI_Assurance_2022	Assuring AI methods for economic policymaking	Anderson Monken, William Ampeh, Flora Haberkorn, Uma Krishnaswamy, and Feras A. Batarseh	*AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI*			371-428	2022	https://www.google.com/books/edition/AI_Assurance/dch6EAAAQBAJ?hl=en&gbpv=1&dq=%22Large+language+models+LLMs+are+a+class+of+language+models+that+use+deep+learning+algorithms+and+are+trained+on+extremely+large+textual+datasets+that+can+be+multiple+terabytes+in+size%22&pg=PA376&printsec=frontcover	The definition for "large language model (LLM)" appears on page 376. This book was edited by Feras A. Batarseh and Laura Freeman.
AI_Fairness_360	Glossary	AI Fairness 360	AI Fairness 360					https://aif360.mybluemix.net/resources#glossary	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
AIMA	Artificial Inelligence: A Modern Approach	Russell, Stuart; Peter Norvig	Pearson				2010	https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf	
Brookings_Institution	The Brookings glossary of AI and emerging technologies	Allen, John R. and Darrell M. West	Brookings Institution				2021	https://www.brookings.edu/blog/techtank/2020/07/13/the-brookings-glossary-of-ai-and-emerging-technologies/	
Brownlee,_Jason_GAN	A Gentle Introduction to Generative Adversarial Networks (GANs)	Brownlee, Jason	Machine Learning Mastery				2019	https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/	
C3.ai_feedback_loop	What Is a Feedback Loop?	C3.ai	C3.ai					https://c3.ai/glossary/features/feedback-loop/	
C3.ai_Model_Training	Model Training	C3.ai	C3.ai Glossary					https://c3.ai/glossary/data-science/model-training/	
COE_AI_Glossary	Artificial Intelligence Glossary		Council of Europe					https://www.coe.int/en/web/artificial-intelligence/glo	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
								ssary	
CSRC	Information Technology Laboratory Computer Security Resource Center Glossary		NIST					https://csrc.nist.gov/glossary	
D. Chandler and R. Munday	A Dictionary of Media and Communication	D. Chandler and R. Munday	Oxford University Press				2011		
DOD_TEVV	Technology Investment Strategy 2015-2018	United States Department of Defense's Test and Evaluation, Verification and Validation (TEVV) Working Group	Technology Investment Strategy 2015-2018				2015	https://defenseinnovationmarketplace.dtic.mil/wp-content/uploads/2018/02/OSD_ATEVV_STRAT_DIST_A_SIGNED.pdf	"trust" definition on page 15; "automation" and "autonomy" definitions on page 2; "validation" and "verification" definitions on page 15
Friedman_et_al_2017	A Survey of Value Sensitive Design	Batya Friedman, David G. Hendry, and	*Foundations and Trends® in Human–Computer*				2017	https://www.nowpublishers.com/article/Details/HCI-015	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
	Methods	Alan Borning	*Interaction*						
Gama_Joao	A Survey on Concept Drift Adaptation							https://repositorio.inesctec.pt/server/api/core/bitstreams/7f101638-0b33-4863-9dd2-cf991f192c9f/content	
GDPR	Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing							https://eur-lex.europa.eu/eli/reg/2016/679/oj	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
	Directive 95/46/EC (General Data Protection Regulation)								
Gustavii,_Ebba	A Swedish Grammar for Word Prediction	Gustavii, Ebba; Eva Pettersson	Uppsala University - Department of Linguistics				2003	https://www.researchgate.net/publication/2838153_A_Swedish_Grammar_for_Word_Prediction/link/00b4951a5f2645ca02000000/download	
humphrey_addressing_2020	Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments: An Urgent Challenge.	Humphrey, Holly J., Dana Levinson, Marc A. Nivet, and Stephen C. Schoenbaum	Academic Medicine	95	12S		2020	https://doi.org/10.1097/ACM.0000000000003679	
Hutson,_Matthew	AI Glossary: Artificial intelligence, in so many	Hutson, Matthew	Science	357	6346	19	2017	https://www.science.org/doi/10.1126/science.357.6346.19	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
	words								
IEEE_Guide_IPA	IEEE Guide for Terms and Concepts in Intelligent Process Automation	IEEE Standards Association	IEEE Guide for Terms and Concepts in Intelligent Process Automation						
IEEE_Soft_Vocab	Systems and software engineering — Vocabulary		ISO/IEC/IEEE 24765				2017	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8016712	
ISO/IEC_TS_5723:2022(en)	ISO/IEC TS 5723:2022(en) Trustworthiness — Vocabulary	ISO/IEC	ISO/IEC				2022	https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en	
Jenna_Burrell	How the machine 'thinks': Understanding opacity in machine learning algorithms	Jenna Burrell	*Big Data & Society*			1-12	2016	https://journals.sagepub.com/doi/pdf/10.1177/2053951715622512	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
NIST_1500	NIST Big Data Interoperability Framework	Wo L. Chang; Nancy Grady	NIST	1			2019	https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions?pub_id=918927	
NIST_AI_RMF_1.0	NIST AI RMF 1.0	NIST	NIST AI RMF 1.0				2023	https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf	Definition 5 for "risk" comes from p. 3 of NIST AI RMF 1.0.
NISTIR_8269_Draft	A Taxonomy and Terminology of Adversarial Machine Learning	Tabassi, Elham;Kevin J. Burns; Michael Hadjimichael; Andres D. Molina-Markham; Julian T. Sexton	Draft NISTIR 8269				2019	https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf	
NSCAI	National Security Commission on Artificial Intelligence: The Final Report	National Security Commission on Artificial Intelligence	National Security Commission on Artificial Intelligence Final Report				2021	https://www.nscai.gov/2021-final-report/	Appendix A: Technical Glossary begins on page 601 of the report (603 of the PDF itself).

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
OECD	Glossary of Statistical Terms	Organisation for Economic Co-operation and Development					2007	https://ec.europa.eu/eurostat/ramon/coded_files/OECD_glossary_stat_terms.pdf / https://stats.oecd.org/glossary/	
privacy-enhancing_technologies	Chapter 5: Privacy-enhancing technologies (PETs)	UK Information Commissioner's Office	DRAFT Anonymisation, pseudonymisation and privacy enhancing technologies guidance				2022	https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf	The definition for "differential privacy" appears on page 30. This document, as accessed on October 27, 2022, was last updated on September 7, 2022.
Public_Health_and_Informatics_MIE_2021	A Preliminary Scoping Study of Federated Learning for the Internet of Medical Things	Arshad Farhad; Sandra I. Woolley; Peter Andras	*Public Health and Informatics: Proceedings of MIE 2021*			504-505	2021	https://www.google.com/books/edition/Public_Health_and_Informatics/81A2EAAAQBAJ?hl=en&gbpv=1&dq=%22Federated+learning+1+2+is+a+learning+	Definition for "federated learning" appears on page 504

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
								model+which+addresses+the+problem+of+data+governance+and+privacy+by+training+algorithms+collaboratively+without+transferring+the+data+to+another+location%22&pg=PA504&printsec=frontcover	
Ranschaert,_Erik	Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks	Ranschaert, Erik R.; Sergey Morozov; Paul R. Algra	Springer				2019	https://link.springer.com/content/pdf/10.1007/978-3-319-94878-2.pdf	
Reznik,_Leon	Introduction I.5 Glossary of Basic Terms	Reznik, Leon	Intelligent Security Systems: How Artificial Intelligence, Machine Learning and Data Science Work for and			xv-xxiv	2022		

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
			Against Computer Security						
Schwartz, R et al.	Towards a Standard for Identifying and Managing Bias in Artificial Intelligence	Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. and Hall, P.	Special Publication (NIST SP 1270), National Institute of Standards and Technology				2022	https://doi.org/10.6028/NIST.SP.1270	
SP1011	Autonomy Levels for Unmanned Systems (ALFUS) Framework	Autonomy Levels for Unmanned Systems Working Group Participants	NIST Special Publication 1011				2008	https://www.nist.gov/system/files/documents/el/isd/ks/NISTSP_1011-I-2-0.pdf	
Steinhardt,_Jacob	Certified Defenses for Data Poisoning Attacks	Steinhardt_Jacob; Pang Wei Koh; Percy Liang	31st Conference on Neural Information Processing Systems				2017	https://proceedings.neurips.cc/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf	
tabassi_adversarial_2019	A Taxonomy and Terminology of Adversarial Machine Learning	Tabassi, Elham, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and	NIST Internal or Interagency Report (NISTIR) 8269 (Draft)				2019	https://doi.org/10.6028/NIST.IR.8269-draft	

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
		Julian Sexton.							
UNODC_Glossary_QA_GLP	Glossary of Terms for Quality Assurance and Good Laboratory Practices	Laboratory and Scientific Section of the United Nations Office on Drugs and Crime	Glossary of Terms for Quality Assurance and Good Laboratory Practices				2009	https://www.unodc.org/documents/scientific/ST_NAR_26_E.pdf	
U.S. Census	What Are Synthetic Data?	U.S. Census					2021	https://www.census.gov/about/what/synthetic-data.html	
Virginia_Dignum_Responsibility_and_Artificial_Intelligence	Responsibility and Artificial Intelligence	Virginia Dignum	*The Oxford Handbook of Ethics of AI*			215-232	2020	https://www.google.com/books/edition/The_Oxford_Handbook_of_Ethics_of_AI/8PQTEAAAQBAJ?hl=en&gbpv=1&dq=%22Understanding+the+values+behind+the+technology+and+	Definition 1 for "values" is taken verbatim from page 221; see the note at the end of the term's row.

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Volume	Issue	Page(s)	Year	URL	Notes
								deciding+on+how+we+want+our+values+to+be+incorporated+in+AI+systems+requires+that+we+are+also+able+to+decide+on+how+and+what+we+want+AI+to+mean+in+our+societies%22&pg=PA221&printsec=frontcover	
yields.io_model_validation	What Is Model Validation?	Eimee V	Yields.io				2020	https://www.yields.io/blog/what-is-model-validation/	Date of publication is February 3, 2020
Žliobaitė_Indrė	A survey on measuring indirect discrimination in machine learning	Žliobaitė, Indrė	CoRR				2018	https://arxiv.org/abs/1511.00148	