EU-U.S. Trade and Technology Council
Working Group 1: Technology Standards
Subgroup on AI Taxonomy & Terminology

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

Delivering on the commitments outlined in the *first edition* of the EU-U.S. Terminology and Taxonomy for Artificial Intelligence (henceforth, the *first edition*), EU and U.S. AI experts from the EU-US Trade and Technology Council (TTC) Working Group 1 (WG1) sub-group on AI Taxonomy and Terminology[1] have developed a *second edition* on the occasion of the Sixth EU-U.S. TTC Ministerial (TTC6).

After a preliminary analysis of external input collected through a call for comments and a series of internal consultations, the WG1 experts added 13 new terms and amended 24 terms from the *first edition* in the development of the *second edition*.

As stated in the EU-U.S. Trade and Technology Council (TTC) 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, the EU and U.S. worked together to prepare an initial draft on AI terminologies and taxonomies. On the occasion of the Fourth EU-U.S. TTC Ministerial (TTC4), members from the EU and U.S. presented the first edition. A total number of 65 terms were defined, with references to key documents from the EU, the United States, and other scholarly sources.

Through the European Commission and the National Institute of Standards and Technology (NIST), external experts had the opportunity to provide their input on the first edition between October 27, 2023 and November 24, 2023. The call for input was published through the European Commission's Directorate General for Communications Networks, Content and Technology (DG CNECT) and through NIST's dedicated "AI Policy Contributions" page.

By soliciting input from a broader network of experts, the document can better reflect diverse perspectives and expertise contributing valuable insights to enhance the overall quality of the work undertaken by the experts of WG1.

The enclosed document reflects an iterative process which both incorporated expert input—to the degree the input corresponded with the previously defined criteria and methodology for the development of the Terminology and Taxonomy—and leveraged the WG1 experts to refine and expand the document.

[1] The WG1 experts Jesse Dunietz (NIST), Gry Hasselbalch (InTouchAI.EU), Irina Orssich (DG CNECT, European Commission), Andrea Renda (CEPS), Reva Schwartz (NIST) and Elham Tabassi (NIST). The work of the experts was supported by Camille Ford (CEPS), Robert Praas (CEPS), Krisstina Rao (InTouchAI.EU), Robert Scholz (DG CNECT) and Freddy Trinh (DG CNECT).

Preliminary Analysis of the Call for Input Responses

Below is a summary of WG1 experts' preliminary analysis of the input received by the European Commission and NIST. The call for input generated 34 total contributions, which were compared against the first edition in three broad categories:

1. Existing terms to which an amendment was proposed, with the source of the updated definition as available. These contributions were grouped by term and cluster, to highlight recurring suggestions and those clusters which received the most input.
2. New terms, their proposed definitions, and their source, as available. These were grouped by term on a near-match basis, to highlight recurring suggestions.
3. Existing terms for which input was not provided.

Several suggestions were cross-cutting or addressed how terms are organized into the taxonomy. WG1 experts analyzed these comments to develop the second edition. Of note, 24 out of 65 terms from the list were not subject to input. The WG1 experts consider this evidence of an emerging consensus on these terms in the transatlantic AI eco-system.

A. Methodology

The methodology used to conduct this analysis is detailed below. All received inputs have been published in the Annex.

1. Methodology: Step-by-Step

Step 1. Initial Processing of Comments

The input was collected and first read for relevance. Those comments which did not provide input on the list of terms or broader cooperation efforts on AI within the TTC were removed. Additionally, those comments that alluded to terms listed as 'pending' in the first edition—for their lack of fixed definition in relevant legislative or institutional documents—were also removed.

Step 2. Initial Review of Comments

The WG1 experts read through the entire set of responses to identify proposed amendments to the current set of definitions and potential new terms to include in the second edition, as provided by the external expert input.

Given the nature of the list, based on their respective areas of expertise, the WG1 experts devoted special attention to proposals backed by authoritative references, and that presented a clear case for improving either the terms themselves or the associated definitions. Where proposals were not accompanied by references, the WG1 experts sought to find supporting literature and official documents to strengthen the case of considering such proposals in the revision of the list.

In reviewing the input, the WG1 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.

The review process prioritized comments that aligned with the primary selection criteria stated in the first edition. These criteria required submitted comments to contribute towards 1) an essential understanding of a risk-based approach to AI and 2) advancing EU-US cooperation on AI.

Finally, some of the submissions provided a valuable contribution by detailing relevant ongoing work by associations and institutions, of which the WG1 experts have taken due notice. However, submissions were considered for incorporation only if they contained concrete proposals to refine or amend the original list of terms.

This process helped identify those clusters and terms which were the subject of most and least input.

Step 3. Specific Input Review and Preliminary Analysis of Findings

The WG1 experts discussed the amendments and terms identified in step 2. They took note of the following:

1. The frequency of input generated for specific existing terms, with close attention to those existing terms which received several proposed amendments and the substance of the suggested amendment.
2. The frequency of input suggesting the same specific new term, with close attention to those recurring terms and their provided definitions.
3. Those terms which were not addressed across the input.

The WG1 experts went over the list of input on existing terms to divide input into three categories. After reviewing the 41 existing terms which were the subject of input, the experts:

1. Revised 21 of the existing terms which were the subject of input, taking into account the suggested amendments as well as the publication of new expert sources and the addition of new expertise in the working group since the publication of the first edition;
2. Considered amendments to 6 existing terms;
3. Did not consider amendments to 14 existing terms, as these did not meet the criteria for revision outlined in the methodology (see: step 2).

The following observations resulted from this review:

1. Certain new terms converge with the discussion on existing terms. As such, the WG1 experts cross-referenced the new term proposals with the existing terms list, and, as relevant, used suggestions of new terms and definitions as amendments to an existing term.
2. Certain new terms did not meet the criteria for consideration or referred to a term which was deemed “pending” in the first edition, and therefore were discarded from the analysis.
3. Certain new terms were deemed too broad or too specific to be included in the scope of the terminology document. These terms were flagged to be reviewed collectively, with the intention of defining the scope of the terminology document more explicitly.

Step 4. *Revisions and Internal Consultations*

The revision process was conducted as follows through a series of internal consultations:

- **Existing terms:** the WG1 experts analyzed the 27 existing terms for which they were considering amendments and provided a revised definition based on the external input and/or cross-referencing the input with internal expertise and authoritative external sources. While the meaning of the input was considered and, where appropriate, reflected in the revision of the terms, direct suggestions for revision of text were not always used directly. Ultimately, 24 terms were amended.
- **New terms:** the WG1 experts reviewed the proposed new terms and assessed their overall relevance to the field and the exercise. Based on the external input and discussions during internal consultations, the WG1 experts selected 9 new terms to be defined. They gathered references to

select or develop a definition of the proposed new terms and organized the terms into the five term "clusters[2]" of the first edition.

- **Pending terms:** the WG1 experts reviewed the pending terms of the first edition to assess whether they should be incorporated into the second edition.

Step 5. ***Development of the Second Edition***

As a final step, the WG1 experts reviewed all the amendments to ensure consensus-based definitions grounded in authoritative sources. The result is the enclosed document.

The WG1 experts thank all external experts for their contributions to this exercise.

[2] The clusters include: AI Lifecycle, Measurement, Technical System Attributes, Governance and Trustworthy.

1. **Cluster: AI Lifecycle**

Term	Definition	Reference
Adversarial machine learning	A field concerned with designing ML algorithms that can resist security challenges, studying the capabilities of attackers, and understanding attack consequences.	Derived from Adversarial_ML and Adversarial_ML_II
Autonomy (Autonomous AI system)	A system's level of independence from human involvement and ability to operate without human intervention. [Different AI systems have different levels of autonomy.] An autonomous system has a set of learning, adaptive and analytical capabilities to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. Autonomous or semi-autonomous AI systems can be characterised as "human-in-the-loop", "human-on-the-loop", or "human-out-of-the-loop" systems depending on their level of meaningful involvement of human beings.	Simplified from DOD_TEVV and AI Act

Big data	An all-encompassing term for large, complex digital data sets whose storage, analysis, management, and processing require similarly complex technological means and substantial computing power. Datasets are sometimes linked together to see how patterns in one domain affect other areas. Data can be structured into fixed fields or unstructured, and are often generated or received at a high rate. The analysis of big datasets, often using AI, can reveal patterns, trends, or underlying relationships that were not previously apparent.	Revision based on call for expert input.
Data augmentation	A technique where the training dataset is increased in size and quality by altering the original training data to create new training examples. to train better machine learning models.	Derived from Data_Augmentation
Data poisoning	A type of security attack where malicious actors modify training data with the aim of corrupting the learned model, thus making the AI system learn something that it should not.	Derived from Adversarial_ML_II
Feature engineering	Feature engineering is the act of extracting features from raw data—i.e., extracting numeric representations of aspects of the data—and transforming them into formats that are suitable for a machine learning model.	Feature_ Engineering_ML
Knowledge representation	The art of formalizing knowledge declaratively, typically for use in a symbolic AI system such as an expert system.	Elsevier_Knowledge_Representation Springer_Knowledge_Representation PsychologyPress_Knowledge_Representation

Lifecyle of an AI system	An AI system lifecycle phases typically involves several phases including: 1) planning and design, 2) data collection and processing, 3) model building and/or adapting existing models to specific tasks, 4) testing, evaluation, verification and validation, 5) deployment, and 6) operation and monitoring. These phases often take place in an iterative manner and are not necessarily sequential.	*OECD Forthcoming*
Loss Function (also called cost function)	The loss function produces a single, overall assessment metric, for training purposes, of an AI system taking any given available decision or action. Typically, the goal of AI system training is to minimize the total loss over some validation set of examples.	Bishop_ML Russell_Norvig_AIMA_4
Machine learning	Machine Learning (ML) is a branch of Artificial Intelligence (AI) that focuses on the development of systems capable of learning from data to perform a task without being explicitly programmed to perform that task. Learning refers to the process of optimizing model parameters through computational techniques such that the model's behaviour is optimized for the training task.	Combination of JRC and ISO_22989.
Natural language processing	The field concerned with machines capable of processing, analysing, and generating human language, either spoken, written or signed.	Revision of own definition based on ISO/IEC in JRC and Hutson_Matthew
Prompt	Prompts are inputs to a generative AI system describing a task that the system should perform or the information it should respond to.	Prompt_Engineering
Prompt engineering	the process of designing and crafting prompts for generative AI models to elicit desired outputs. It involves understanding the capabilities of the model and tailoring the prompt to effectively guide the model towards generating relevant, informative, and creative outputs.	Prompt_Engineering

Reinforcement learning	Reinforcement learning (RL) is a subset of machine learning that allows an artificial system (sometimes referred to as an agent) in a given environment to optimize its behaviour. Agents learn from feedback signals received as a result of their actions, such as rewards or punishments, with the aim of maximizing the received reward. Such signals are computed based on a given reward function, which constitutes an abstract representation of the system's goal. The goal could be, for example, to earn a high video game score or to minimize idle worker time in a factory.	Derived from Dayan_Niv_RL; Sutton_RL; HuggingFace_RL ; and Dayan_Watkins_RL
Synthetic data	Synthetic data is data artificially generated by a computational process rather than being captured by sensory apparatus or manually created by humans. Synthetic data is often produced by a model trained to reproduce the characteristics and structure of its training data, aiming for similar distribution.	Revision of own definition based on EDPS_SD
Training data	Data used for training an AI system through fitting its learnable parameters, including the weights of a neural network.	Derived from AI Act

2. Cluster: Measurement

Term	Definition	Reference
(AI) accuracy	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. The concept of accuracy is often used to evaluate the predictive capability of the AI model.	Revision of combination based on EU HLEG/ALTAI and OECD.

3. Cluster: Technical System Attributes

Term	Definition	Reference
AI System	An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.	OECD_AI_System
Adaptive learning (adaptiveness)	**Adaptiveness** is the characteristic of some AI systems of being able to change their behaviour while in use based on interactions with input and data. [Adaptation may entail a change in the weights of the model or a change in the internal structure of the model itself.] [Examples include a speech recognition system that adapts to an individual's voice or a personalised music recommender system.] The new behaviour of the adapted system may produce different results than the previous system for the same inputs.	OECD_AI_System
Expert system	Automated systems encoded with knowledge of human experts, typically through knowledge representation techniques. Focused on narrow tasks and with automated decision-making based on "if-then" rules.	Crevier
Federated learning	Federated learning is an approach to machine learning which addresses problems of data governance and privacy by training algorithms collaboratively without transferring the data to a central location. Each federated device trains on data locally and shares its local model parameters instead of sharing the training data. Different federated learning systems have different topologies that involve different ways of sharing parameters.	Revision of own definition based on combination of EDPS_FL and Public_Health_and_Informatics_MIE_2021

Human values for AI	AI systems are not value neutral. Values are idealised qualities or conditions in the world that people find good. The design of human values in AI systems implies negotiation of different values and values-systems of meaning making, and it requires decisions regarding ethical principles, governance, policies and incentives. Designing AI with human values will necessitate awareness of the social and economic interests behind AI systems as well as respect for cultural diversity.	Revision of own definition based on EU and U.S. values and Brey
Human-centric AI	Human-Centric AI (or "human-centered AI") is an approach to the design, deployment and use of AI systems that considers them as components of socio-technical environments in which humans assume meaningful agency. The Human-Centric Approach to AI prioritizes enhancing human capabilities rather than replacing them. The approach is promoted in policy, research and engineering with the aim to develop AI systems as tools to serve human beings and to increase human and environmental well-being by promoting human rights, the rule of law, democratic values and sustainable development.	Revision of first edition based on Hasselbalch; HLEG AI, Ethics Guidelines for Trustworthy AI; Shneiderman_Trustworthy
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. Most LLMs can generate text, although some can only form a compressed representation of an input useful for tasks such as classification or question answering.	Revision of own edition based on AI_Assurance_2022
Model	A core component of an AI system used to make inferences from inputs in order to produce outputs. A model characterizes an input-to-output transformation intended to perform a core computational task of the AI system (e.g., classifying an image, predicting the next word for a sequence, or selecting a robot's next action given its state and goals).	Own definition based on OECD_AI_System and AI_Fairness_360

Neural network	A neural network consists of one or more layers of neurons connected by weighted links with adjustable weights. A neural network receives input data and produces an output by passing it through the network, with each neuron performing a simple computation. Whereas some neural networks are intended to simulate the functioning of biological neurons in the nervous system, most neural networks in artificial intelligence are engineering tools that draw only loose inspiration from biology.	Revision from first edition, own definition based on ISO/IEC in JRC and Ranschaert,_Erik

4. Cluster: Governance

Term	Definition	Reference
Auditability of an AI system	Auditability refers to the ability of an AI system to undergo assessment of its algorithms, data and design processes, in particular to determine whether the system is working as intended. Auditability 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.	Revision of own definition based on EU HLEG/ALTAI

5. Cluster: Trustworthy

Term	Definition	Reference
Bias	An AI system's differential treatment of different groups, which may arise from implicit systems of meaning, norms and values. See also "harmful bias" and "discrimination."	Revised definition based on humphrey_addressing_2020
Harmful Bias	Biases of an AI system that create negative impacts such as unfair or discriminatory outcomes (see also "Discrimination"). Different types of harmful bias emerge due to many factors, including but not limited to human or system decisions and processes across the AI lifecycle; pre-existing cultural and social bias in training data; technical limitations (such as non-representative or limited design specifications and data); or use in unanticipated contexts. Measures can be put in place to mitigate and detect bias.	Revised definition based on humphrey_addressing_2020
Confabulation (also known as Hallucination)	When generative AI systems generate inaccurate or false responses that can appear plausible to the user. Confabulation can, e.g., be the invention of erroneous historical or biographical information. AI confabulation is the result of statistical prediction, repetition of training data or patterns.	Own definition based on IEEE_Hallucination
Data Leakage	[in the context of AI] Data leakage is the introduction of information a system will be expected to infer into the data it is trained on, which should not be legitimately available to learn from. This results in a high-performing model while evaluating the model on the test set during model development, but poor performance during deployment when evaluated on new data sets.	Own definition based on ACM_Leakage
Deep fake	AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful.	AI Act

Discrimination	Differential treatment of individuals based on factors such as their ethnicity, culture or religion. Discrimination can be a result of institutional and individual biases that are embedded in processes across the AI lifecycle, e.g. cultural and social biases held by AI actors and organisations, or represented in the data of AI systems. Discrimination can also be the result of technical limitations in hardware or software, or of the use of an AI system that, due to its context of application, does not treat all groups equally. As many forms of biases are systemic and implicit, they are not easily controlled or mitigated and require specific governance and other similar approaches.	Based on Boyd_Critical_Questions and UN_General_Assembly
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, affecting their behaviour. A typical example is to change some pixels in a picture before uploading, so that an image recognition system fails to classify the result correctly. Evasion can also be used during deployment.	Own definition based HLEG_ALTAI_Assessment_List and Adversarial_ML_II
Opacity	When one or more features of an AI system, such as processes, the provenance of datasets, functions, output or behaviour are unavailable or incomprehensible to all stakeholders – usually an antonym for transparency.	Revision of own definition based on call for expert input and Jenna_Burrell

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, even with good intentions, AI systems can cause unintentional harm. 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. Trustworthy AI is based on respect for human rights and democratic values.	Revision of own definition based on NIST_AI_RMF_1.0 and EU HLEG/ALTAI

Annex A. Call for Expert Input

Please find the external expert input here. Please note the input has been stripped of personally identifying information.

Annex B. Overview: New and Amended Terms

New Terms (13)

Cluster 1: AI Lifecycle
1. Data augmentation
2. Feature engineering
3. Knowledge representation
4. Lifecyle of an AI system*
5. Loss function (or cost function)
6. Prompt
7. Prompt engineering
8. Training data*

Cluster 3: Technical System Attributes

9. AI System*
10. Expert system

Cluster 5: Trustworthy

11. Confabulation (sometimes called hallucination)
12. Data Leakage
13. Deep Fake*

Amended Terms (24):

Cluster 1: AI Lifecycle
1. Adversarial machine learning (adversarial attack)
2. Autonomy (Autonomous AI system)
3. Big data
4. Data poisoning
5. Deep learning
6. Machine learning
7. Natural language processing
8. Reinforcement learning
9. Synthetic data

Cluster 2: Measurement

10. (AI) accuracy

Cluster 3: Technical System Attributes

11. Adaptive learning
12. Federated learning
13. Human values for AI
14. Human-centric AI
15. Large language model (LLM)
16. Model
17. Neural network

Cluster 4: Governance

18. Auditability of an AI system

Cluster 5: Trustworthy

19. Bias
20. Harmful bias
21. Discrimination
22. Evasion
23. Opacity
24. Trustworthy AI

Annex C. References

ID	Title of article, chapter, or page	Author(s) and/or Editor(s)	Publication or website	Vol	Issue	Page(s)	Year	URL	Notes
Adversarial_ML	Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations	Elham Tabassi Kevin J. Burns Michael Hadjimichael Andres D. Molina-Markham Julian T. Sexton	U.S. National Institute of Standards and Technology (NIST)				2019	https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf	
Adversarial_ML_II	Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations	Apostol Vassilev Alina Oprea Alie Fordyce Hyrum Anderson	U.S. National Institute of Standards and Technology (NIST)				2023	https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf	
ACM_Leakage	Leakage in Data Mining: Formulation,	Shachar Kaufman, Saharon Rosset,	*ACM Transactions on*	6	4		2011	https://www.cs.umb.edu/~ding/history/470_670_fall	

	Detection, and Avoidance	Claudia Perlich	*Knowledge Discovery from Data*					_2011/papers/cs670_Tran_PreferredPaper_LeakingInDataMining.pdf	
AI Act	Artificial Intelligence Act	European Commission, European Parliament, Council of the European Union					2024	https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf	
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	The definition for "large language model (LLM)" appears on page 376. This book was edited by Feras A. Batarseh and Laura Freeman.

								+terabytes+in+size%22&pg=PA376&printsec=frontcover	
AI_Fairness_360	Glossary	AI Fairness 360	AI Fairness 360					https://aif360.mybluemix.net/resources#glossary	
Bishop_ML	Pattern Recognition and Machine Learning	Bishop, Christopher M.	Springer				2006	Link	
Boyd_Critical_Questions	Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon	Danah Boyd, Kate Crawford	*Információs Társadalom*			33	2012	https://www.researchgate.net/publication/281748849_Critical_questions_for_big_data_Provocations_for_a_cultural_technological_and_scholarly_phenomenon	
Brey	Values in technology and disclosive ethics	Brey, P.	L. Floridi (ed.) The Cambridge Handbook of Information and			41–58	2010	https://research.utwente.nl/en/publications/values-in-technology-and-disclosive-computer-ethics	

			Computer Ethics, Cambridge University Press						
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/	
Crevier	The Tumultuous History of the Search for Artificial Intelligence	Crevier, D	Basic Books				1993	AI: the tumultuous history of the search for artificial intelligence | Guide books | ACM Digital Library	
Data_Augmentation	A survey on Image Data Augmentation for Deep Learning	Shorten, C., Khoshgoftaar, T.M.	Journal of Big Data	**6**	**60**		2019	Link	

Dayan_Niv_RL	Reinforcement Learning: The Good, The Bad and The Ugly	Dayan, Peter and Niv, Yael	Elsevier	**18**		**185-196**	2008	https://www.princeton.edu/~yael/Publications/DayanNiv2008.pdf	
Dayan_Watkins_RL	Reinforcement learning: A computational perspective.	Dayan, P., & Watkins, C. J.	Encyclopedia of Cognitive Science				2006	https://www.researchgate.net/publication/230504820_Reinforcement_Learning_A_Computational_Perspective	
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
DL_1	Representation Learning: A Review and New Perspectives	Bengio, Y; Courville, A; Vincent, P.	IEEE Transactions on Pattern Analysis and				2013	https://arxiv.org/pdf/1206.5538.pdf	

			Machine Intelligence						
DL_2	Deep Learning in Neural Networks: An Overview	Schmidhuber, J.	Neural Networks				2015	https://arxiv.org/abs/1404.7828	
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	
Elsevier_Knowledge_Representation	Handbook of Knowledge Representation	Frank van Harmelen , Vladimir Lifschitz , Bruce Porter	Elsevier				2008	Link	
EU HLEG/ALTAI	Assessment List for Trustworthy	European Union High Level Expert					2020	https://digital-strategy.ec.europa.eu/en/library/asse	

	Artificial Intelligence (ALTAI) for self-assessment	Group on Artificial Intelligence						ssment-list-trustworthy-artificial-intelligence-altai-self-assessment	
Feature_ Engineering_ ML	Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists	Alice Zheng, Amanda Casari	O'Reilly Media, Inc				2018	Link	
Hasselbalch		Hasselbalch, 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	"Socio-technical systems" definition based on STS literature: Hughes, T.P. (1987) The evolution of large technological systems. In W.E. Bijker, T.P. Hughes, T.Pinch

									(eds.) The social construction of technological systems. MIT Press; Hughes, T.P. (1983) Networks of power:Electrification in Western society 1880–1930. The John Hopkins University Press; Bijker, W.E., Hughes, T.P., Pinch, T. 22 (eds.) (1987) The social construction of technological

									systems. MIT Press; Misa, T.J. (1988) How Machines Make History, and How Historians (And Others) Help Them to Do So. Science, Technology, and Human Values, 13(3/4) (Summer – Autumn, 1988), 308–331; Misa, T.J. (1992) Theories of Technological Change:Parameter s and Purposes. Science,

									Technology, and Human Values, 17(1) (Winter, 1992); Misa, T.J. (2009) Findings follow framings: navigating the empirical turn. Synthese, 168. Bijker, W.E., Law, J. (eds.) (1992) Shaping technology/building society: Studies in sociotechnical 23 change. MIT Press; Edwards, P. (2002

HLEG/ALTAI_Assessment_List	The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment	High-level expert group on AI set up by the European Commission	*European Commission*				2020	https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342	
HLEG AI, Ethics Guidelines for Trustworthy AI	Ethics guidelines for trustworthy AI	European Commission					2019	https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai	
HuggingFace_RL	Illustrating Reinforcement Learning from Human Feedback (RLHF)	Lambert, N; Castricato, L; Havrilla, A; von Werra, L	Hugging Face				2022	https://huggingface.co/blog/rlhf	
Humphrey_addressing_2020	Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning	Humphrey, Holly J., Dana Levinson, Marc A. Nivet, and Stephen C. Schoenbaum	Academic Medicine	95	12S		2020	https://doi.org/10.1097/ACM.0000000000003679	

	Environments: An Urgent Challenge.								
Hutson_Matthew	AI Glossary: Artificial intelligence, in so many words	Hutson, Matthew	Science	357	6346	19	2017	https://www.science.org/doi/10.1126/science.357.6346.19	
IEEE_Hallucinations	AI Hallucinations: A Misnomer Worth Clarifying	Negar Maleki Balaji Padmanabhan Kaushik Dutta	*IEEE*				2024	https://arxiv.org/html/2401.06796v1	
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	

ISO_22989	Artificial intelligence concepts and terminology	ISO/IEC	ISO/IEC				2022	https://www.iso.org/standard/74296.html	
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	
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	
Nature_DL	Deep learning	LeCun, Y., Bengio, Y. & Hinton, G	*Nature*	**521**		436–444	2015	Link	
NIST_AI_RM	NIST AI RMF	NIST	NIST AI				2023	https://nvlpubs.nist.gov/nistpubs/ai/	Definition 5 for "risk" comes from p. 3 of NIST

F_1.0	1.0		RMF 1.0					NIST.AI.100-1.pdf	AI RMF 1.0.
OECD	Glossary of Statistical Terms	Organisation for Economic Co-operation and Development	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/	
OECD_AI_Society	Artificial Intelligence in Society	Organisation for Economic Co-operation and Development (OECD)	Organisation for Economic Co-operation and Development (OECD)					https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en#:~:text=The%20AI%20system%20lifecycle&text=The%20design%2C%20data%20and%20models,operation%20and%20monitoring%20(Figure%201.5.	

OECD_AI System	Explanatory memorandum on the updated OECD definition of an AI system	Organisation for Economic Co-operation and Development	Organisation for Economic Co-operation and Development	**8**			2024	Link	
Prompt_Engineering	Prompt Engineering in Large Language Models.	Marvin, Ggaliwango & Hellen Raudha, Nakayiza & Jjingo, Daudi & Nakatumba-Nabende, Joyce.	Springer Nature Singapore			**387-402**	2024	URL	
PsychologyPress_Knowledge_Representation	Knowledge Representation	Arthur B. Markman	Psychology Press				2013	Link	
Public_Health_and_Informatics_MIE_2021	A Preliminary Scoping Study of Federated Learning for the Internet of	Arshad Farhad; Sandra I. Woolley; Peter Andras	*Public Health and Informatics:*			504-505	2021	URL	Definition for "federated learning" appears on page 504

	Medical Things		*Proceedings of MIE 2021*						
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	
Russell_Norvig_AIMA_4	Artificial Intelligence: A Modern Approach	Russell, Stuart and Norvig, Peter.	Pearson Education	**4**			2021	https://aima.cs.berkeley.edu/	
Shneiderman_Guidelines		Shneiderman, B.						https://dl.acm.org/doi/abs/10.1145/3419764	Human-centric AI
Springer_Knowledge_Representation	Knowledge Representation	Grega Jakus, Veljko Milutinović, Sanida Omerović, Sašo Tomažič	Springer			47–62	2013	Link	
Sutton_RL	Reinforcement Learning: An Introduction	Sutton, R.S. and Barto A.G.	MIT Press				2020	http://incompleteideas.net/book/RLbook2020.pdf	

UN_General_ Assembly	Human rights and cultural diversity	General Assembly United Nations	*Report of the Third Committee A/70/489/ Add.2*			5	2015	https://www.refworld.org/legal/resolution/unga/2016/en/109325	