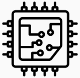
AI standardization in support of the AI Act: what role for academia and research?

Lauriane Aufrant

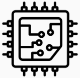
ERCIM seminar

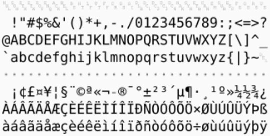
22 March 2024

What is the common point between...



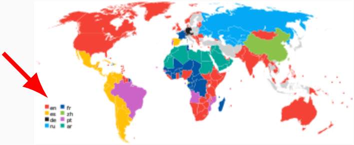
What is the common point between...





```
 !"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~

 ¡¢£¤¥¦§¨©ª«¬-®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ
```



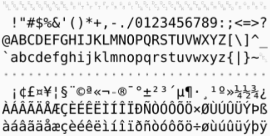
What is the common point between...

THE C PROGRAMMING LANGUAGE

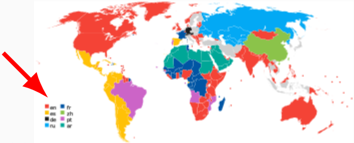
ISO/IEC 9899

ISO/IEC 11172-3

ISO/IEC 8859-1

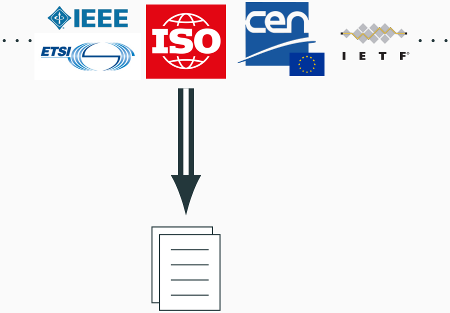


```
 !"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
 ¡¢£¤¥¦§¨©ª«¬-®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ
```

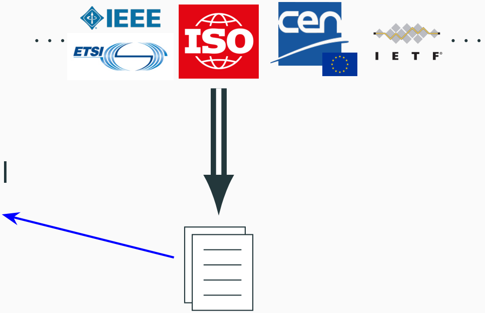


ISO 639-1

Standards

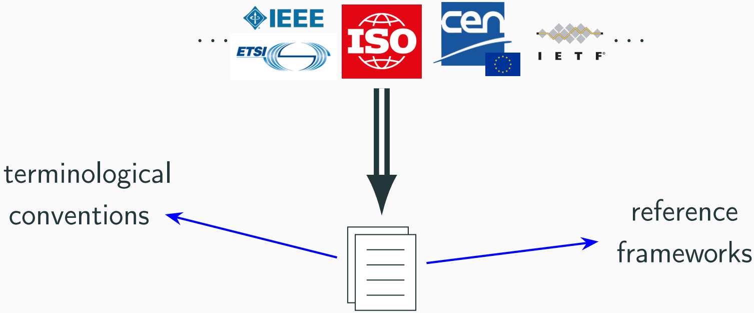


Standards



terminological conventions

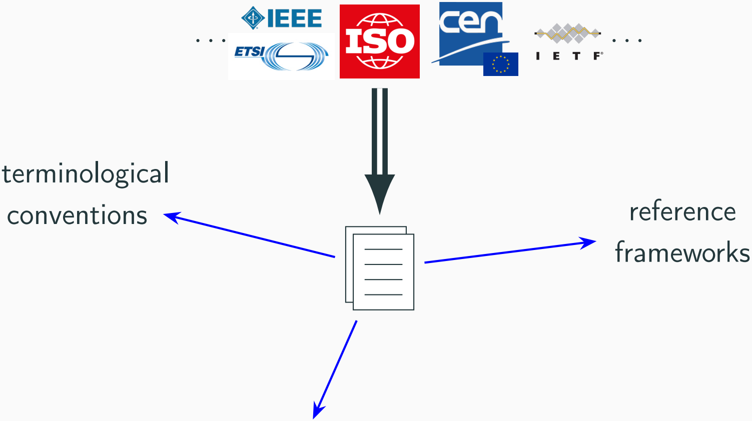
Standards



terminological conventions

reference frameworks

Standards



terminological conventions

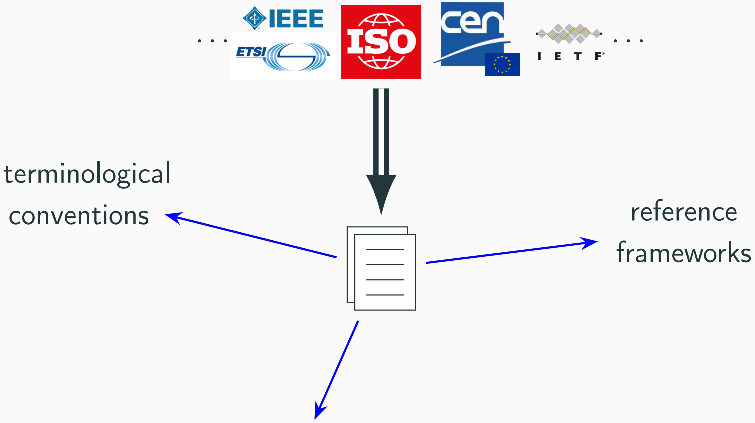
reference frameworks

technical specifications

↪ for data, systems, procedures...

↪ quality metrics, interoperability (APIs, protocols...)

Standards



terminological conventions

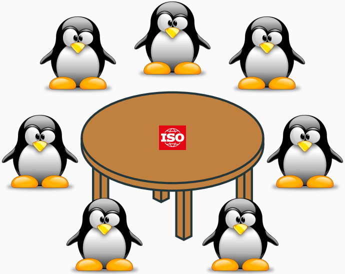
reference frameworks

technical specifications

↪ for data, systems, procedures...

↪ quality metrics, interoperability (APIs, protocols...)

compliance test suites
certification


ISO

ISO

ISO

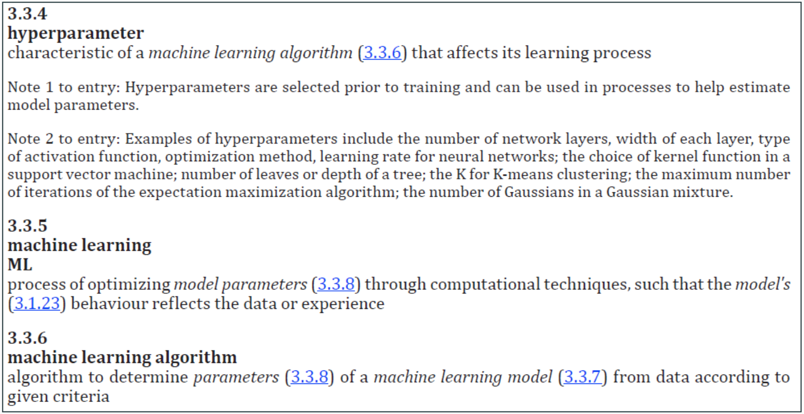
AI standards: what, why?

Formalizing existing ideas

- Interoperability, with consistent data and annotations
- Reproducible and comparable evaluation, with fully specified metrics
- Good practices in annotation, evaluation...
- Ethical guidelines, data statements & model cards...
- ...

  ↪ starting with consistent terminology!

ISO/IEC 22989:2022 *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*



3.3.4
hyperparameter
characteristic of a *machine learning algorithm* (3.3.6) that affects its learning process

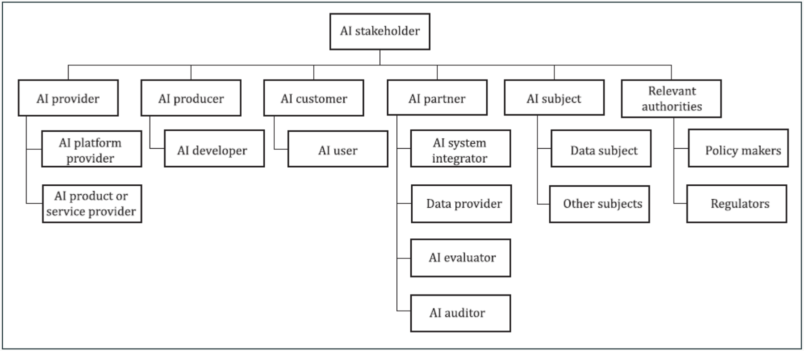
Note 1 to entry: Hyperparameters are selected prior to training and can be used in processes to help estimate model parameters.

Note 2 to entry: Examples of hyperparameters include the number of network layers, width of each layer, type of activation function, optimization method, learning rate for neural networks; the choice of kernel function in a support vector machine; number of leaves or depth of a tree; the K for K-means clustering; the maximum number of iterations of the expectation maximization algorithm; the number of Gaussians in a Gaussian mixture.

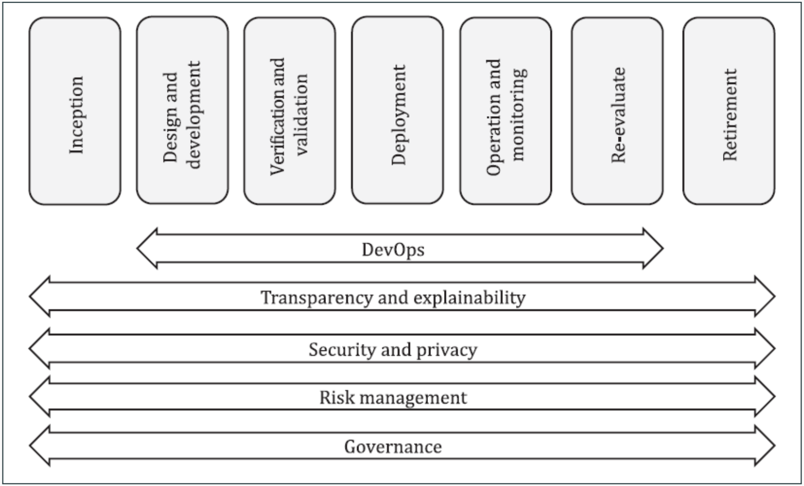
3.3.5
machine learning
ML
process of optimizing *model parameters* (3.3.8) through computational techniques, such that the *model's* (3.1.23) behaviour reflects the data or experience

3.3.6
machine learning algorithm
algorithm to determine *parameters* (3.3.8) of a *machine learning model* (3.3.7) from data according to given criteria

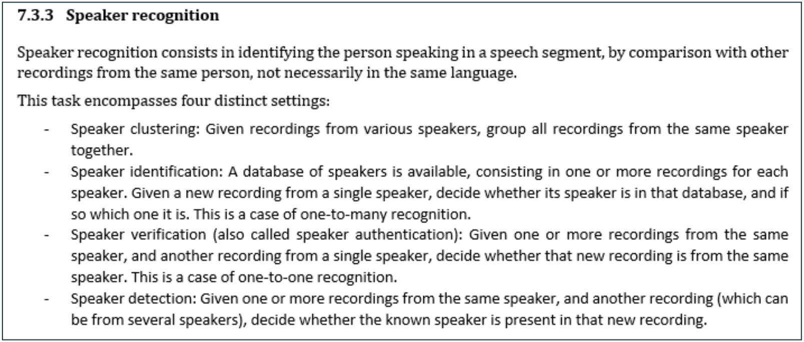
ISO/IEC 22989:2022 *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*



ISO/IEC 22989:2022 *Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*



ISO/IEC CD TR 23281 *Artificial intelligence – Overview of AI tasks and functionalities related to NLP* (draft)



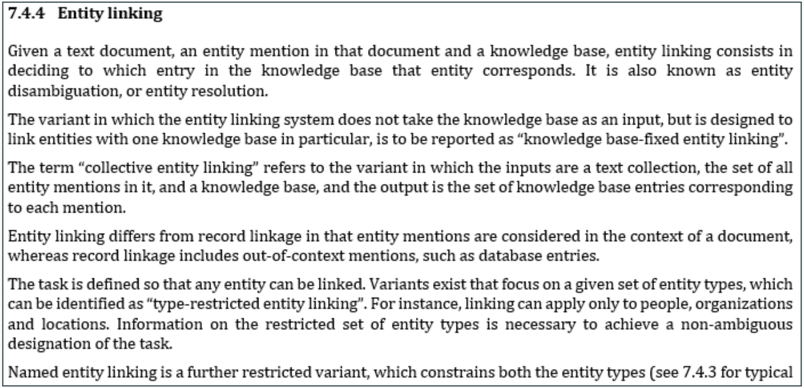
7.3.3 Speaker recognition

Speaker recognition consists in identifying the person speaking in a speech segment, by comparison with other recordings from the same person, not necessarily in the same language.

This task encompasses four distinct settings:

- Speaker clustering: Given recordings from various speakers, group all recordings from the same speaker together.
- Speaker identification: A database of speakers is available, consisting in one or more recordings for each speaker. Given a new recording from a single speaker, decide whether its speaker is in that database, and if so which one it is. This is a case of one-to-many recognition.
- Speaker verification (also called speaker authentication): Given one or more recordings from the same speaker, and another recording from a single speaker, decide whether that new recording is from the same speaker. This is a case of one-to-one recognition.
- Speaker detection: Given one or more recordings from the same speaker, and another recording (which can be from several speakers), decide whether the known speaker is present in that new recording.

ISO/IEC CD TR 23281 *Artificial intelligence – Overview of AI tasks and functionalities related to NLP* (draft)



7.4.4 Entity linking

Given a text document, an entity mention in that document and a knowledge base, entity linking consists in deciding to which entry in the knowledge base that entity corresponds. It is also known as entity disambiguation, or entity resolution.

The variant in which the entity linking system does not take the knowledge base as an input, but is designed to link entities with one knowledge base in particular, is to be reported as "knowledge base-fixed entity linking".

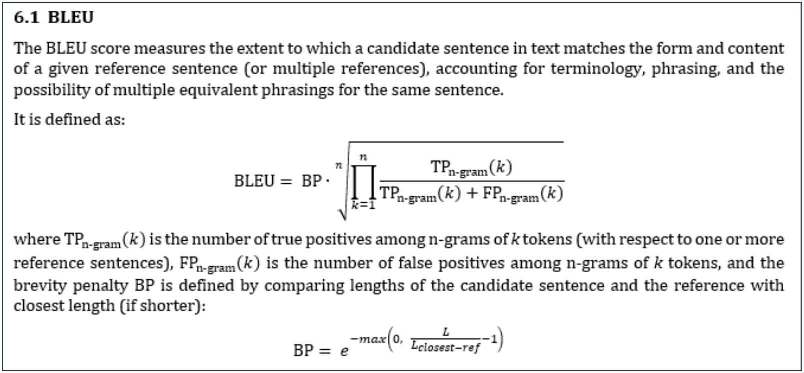
The term "collective entity linking" refers to the variant in which the inputs are a text collection, the set of all entity mentions in it, and a knowledge base, and the output is the set of knowledge base entries corresponding to each mention.

Entity linking differs from record linkage in that entity mentions are considered in the context of a document, whereas record linkage includes out-of-context mentions, such as database entries.

The task is defined so that any entity can be linked. Variants exist that focus on a given set of entity types, which can be identified as "type-restricted entity linking". For instance, linking can apply only to people, organizations and locations. Information on the restricted set of entity types is necessary to achieve a non-ambiguous designation of the task.

Named entity linking is a further restricted variant, which constrains both the entity types (see 7.4.3 for typical

ISO/IEC AWI 23282 *Artificial intelligence – Evaluation methods for accurate NLP systems* (draft)



6.1 BLEU

The BLEU score measures the extent to which a candidate sentence in text matches the form and content of a given reference sentence (or multiple references), accounting for terminology, phrasing, and the possibility of multiple equivalent phrasings for the same sentence.

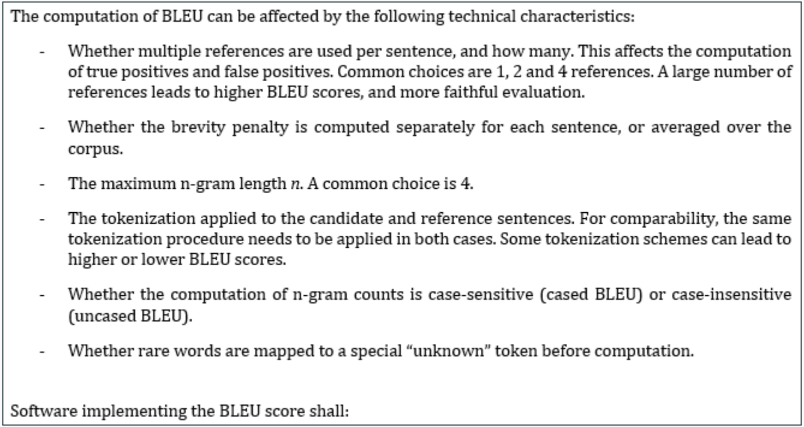
It is defined as:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{\mathrm{TP}_{\text{n-gram}}(k)}{\mathrm{TP}_{\text{n-gram}}(k) + \mathrm{FP}_{\text{n-gram}}(k)}}$$

where $\mathrm{TP}_{\text{n-gram}}(k)$ is the number of true positives among n-grams of k tokens (with respect to one or more reference sentences), $\mathrm{FP}_{\text{n-gram}}(k)$ is the number of false positives among n-grams of k tokens, and the brevity penalty BP is defined by comparing lengths of the candidate sentence and the reference with closest length (if shorter):

$$\mathrm{BP} = e^{-max\left(0,\ \frac{L}{L_{closest-ref}}-1\right)}$$

ISO/IEC AWI 23282 *Artificial intelligence – Evaluation methods for accurate NLP systems* (draft)

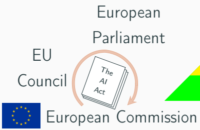


The computation of BLEU can be affected by the following technical characteristics:

- Whether multiple references are used per sentence, and how many. This affects the computation of true positives and false positives. Common choices are 1, 2 and 4 references. A large number of references leads to higher BLEU scores, and more faithful evaluation.
- Whether the brevity penalty is computed separately for each sentence, or averaged over the corpus.
- The maximum n-gram length *n*. A common choice is 4.
- The tokenization applied to the candidate and reference sentences. For comparability, the same tokenization procedure needs to be applied in both cases. Some tokenization schemes can lead to higher or lower BLEU scores.
- Whether the computation of n-gram counts is case-sensitive (cased BLEU) or case-insensitive (uncased BLEU).
- Whether rare words are mapped to a special “unknown” token before computation.

Software implementing the BLEU score shall:

Standards & the EU AI Act

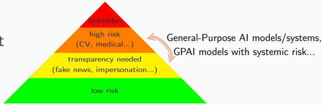


European Parliament

EU Council

The AI Act

European Commission



forbidden

high risk (CV, medical...)

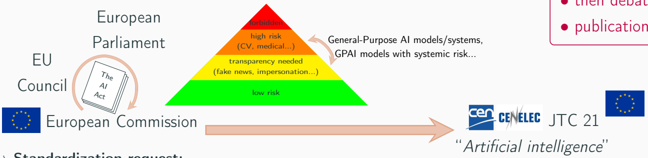
transparency needed (fake news, impersonation...)

low risk

General-Purpose AI models/systems, GPAI models with systemic risk...

Target dates for main AI Act standards:

- core drafting = until mid-2024
- then debating...
- publication = April 2025



European Parliament

EU Council

The AI Act

European Commission

forbidden

high risk (CV, medical...)

transparency needed (fake news, impersonation...)

low risk

General-Purpose AI models/systems, GPAI models with systemic risk...

JTC 21

"*Artificial intelligence*"

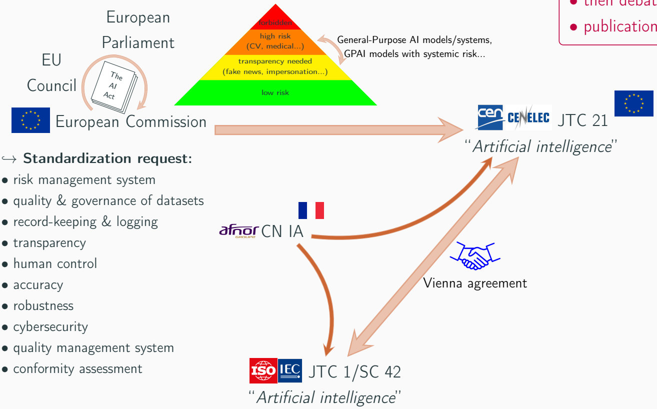
↪ **Standardization request:**

- risk management system
- quality & governance of datasets
- record-keeping & logging
- transparency
- human control
- accuracy
- robustness
- cybersecurity
- quality management system
- conformity assessment

AI Act Article 40:
harmonized standard =
presumption of conformity

Target dates for main AI Act standards:

- core drafting = until mid-2024
- then debating...
- publication = April 2025



European Parliament
EU Council
The AI Act
European Commission
forbidden
high risk (CV, medical...)
transparency needed (fake news, impersonation...)
low risk
General-Purpose AI models/systems, GPAI models with systemic risk...
JTC 21
"*Artificial intelligence*"
afnor CN IA
Vienna agreement
JTC 1/SC 42
"*Artificial intelligence*"

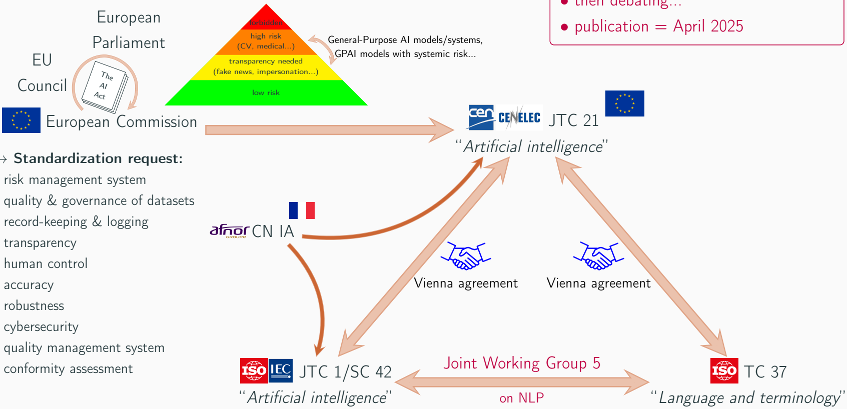
↪ **Standardization request:**

- risk management system
- quality & governance of datasets
- record-keeping & logging
- transparency
- human control
- accuracy
- robustness
- cybersecurity
- quality management system
- conformity assessment

AI Act Article 40:
harmonized standard =
presumption of conformity

Target dates for main AI Act standards:

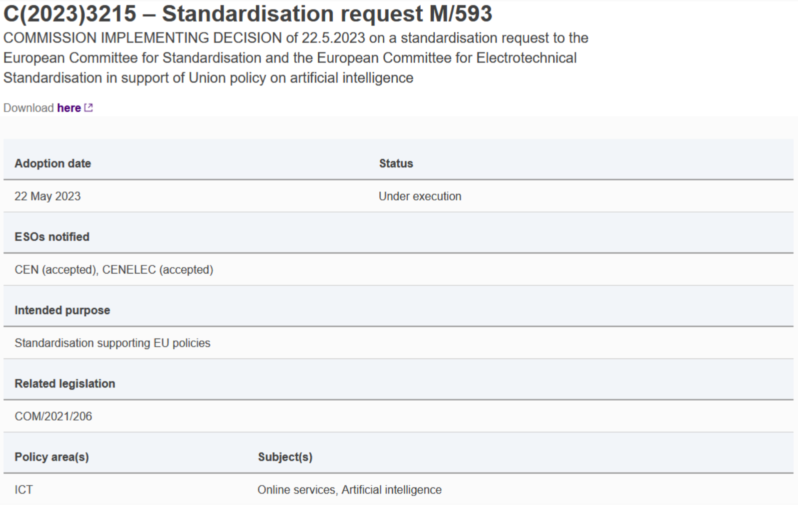
- core drafting = until mid-2024
- then debating...
- publication = April 2025



↪ **Standardization request:**

- risk management system
- quality & governance of datasets
- record-keeping & logging
- transparency
- human control
- accuracy
- robustness
- cybersecurity
- quality management system
- conformity assessment

AI Act Article 40:
harmonized standard =
presumption of conformity



C(2023)3215 – Standardisation request M/593

COMMISSION IMPLEMENTING DECISION of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence

Download **here**

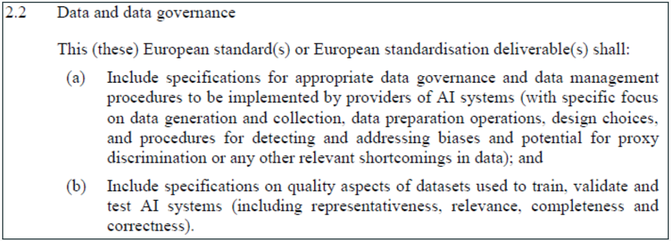
Adoption date	Status
22 May 2023	Under execution

ESOs notified
CEN (accepted), CENELEC (accepted)

Intended purpose
Standardisation supporting EU policies

Related legislation
COM/2021/206

Policy area(s)	Subject(s)
ICT	Online services, Artificial intelligence

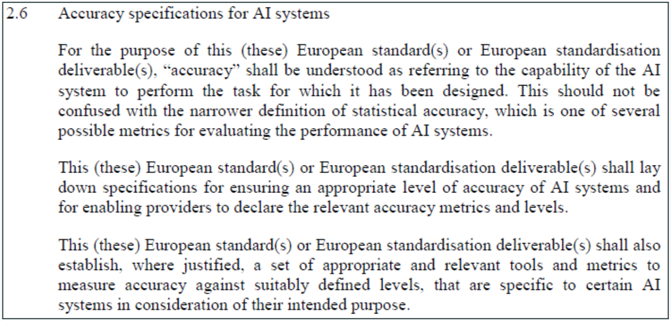


2.2 Data and data governance

This (these) European standard(s) or European standardisation deliverable(s) shall:

(a) Include specifications for appropriate data governance and data management procedures to be implemented by providers of AI systems (with specific focus on data generation and collection, data preparation operations, design choices, and procedures for detecting and addressing biases and potential for proxy discrimination or any other relevant shortcomings in data); and

(b) Include specifications on quality aspects of datasets used to train, validate and test AI systems (including representativeness, relevance, completeness and correctness).

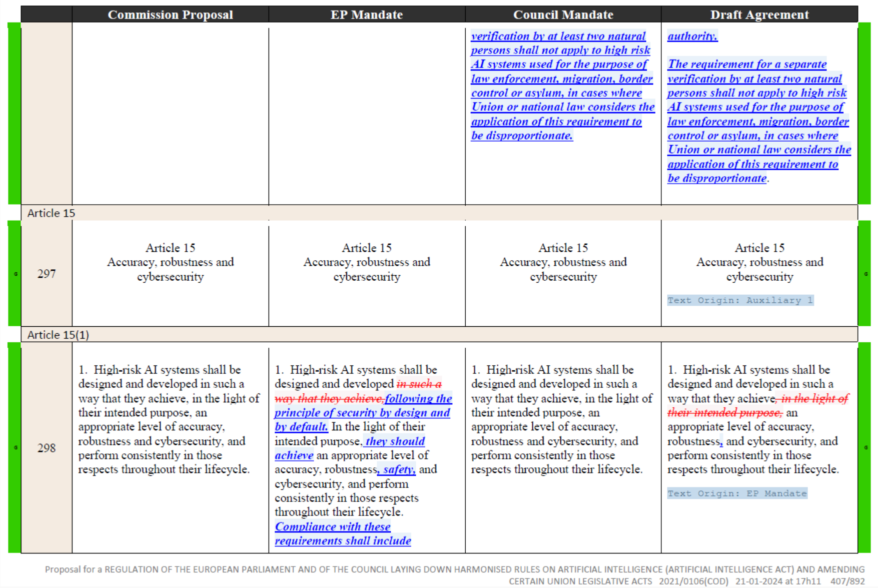


2.6 Accuracy specifications for AI systems

For the purpose of this (these) European standard(s) or European standardisation deliverable(s), “accuracy” shall be understood as referring to the capability of the AI system to perform the task for which it has been designed. This should not be confused with the narrower definition of statistical accuracy, which is one of several possible metrics for evaluating the performance of AI systems.

This (these) European standard(s) or European standardisation deliverable(s) shall lay down specifications for ensuring an appropriate level of accuracy of AI systems and for enabling providers to declare the relevant accuracy metrics and levels.

This (these) European standard(s) or European standardisation deliverable(s) shall also establish, where justified, a set of appropriate and relevant tools and metrics to measure accuracy against suitably defined levels, that are specific to certain AI systems in consideration of their intended purpose.



	Commission Proposal	EP Mandate	Council Mandate	Draft Agreement
			verification by at least two natural persons shall not apply to high risk AI systems used for the purpose of law enforcement, migration, border control or asylum, in cases where Union or national law considers the application of this requirement to be disproportionate.	*authority.* *The requirement for a separate verification by at least two natural persons shall not apply to high risk AI systems used for the purpose of law enforcement, migration, border control or asylum, in cases where Union or national law considers the application of this requirement to be disproportionate*.
Article 15				
297	Article 15 Accuracy, robustness and cybersecurity	Article 15 Accuracy, robustness and cybersecurity	Article 15 Accuracy, robustness and cybersecurity	Article 15 Accuracy, robustness and cybersecurity Text Origin: Auxiliary 1
Article 15(1)				
298	1. High-risk AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle.	1. High-risk AI systems shall be designed and developed ~~in such a way that they achieve,~~*following the principle of security by design and by default.* In the light of their intended purpose, *they should achieve* an appropriate level of accuracy, robustness*, safety,* and cybersecurity, and perform consistently in those respects throughout their lifecycle. *Compliance with these requirements shall include*	1. High-risk AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle.	1. High-risk AI systems shall be designed and developed in such a way that they achieve~~, in the light of their intended purpose,~~ an appropriate level of accuracy, robustness*,* and cybersecurity, and perform consistently in those respects throughout their lifecycle. Text Origin: EP Mandate

Article 15 – Accuracy, robustness and cybersecurity

1. High-risk AI systems shall be designed and developed in such a way that they **achieve an appropriate level of accuracy, robustness, and cybersecurity**, and perform consistently in those respects throughout their lifecycle.

1a. To address the technical aspects of how to **measure the appropriate levels of accuracy and robustness** set out in paragraph 1 of this Article and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholder and organisations such as metrology and benchmarking authorities, encourage as appropriate, the **development of benchmarks and measurement methodologies**.

2. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use. [...]

Article 15 – Accuracy, robustness and cybersecurity

3. High-risk AI systems shall be as resilient as possible regarding **errors, faults or inconsistencies** that may occur within the system or the environment [...].

4. High-risk AI systems shall be resilient as regards to attempts by unauthorised third parties to alter their use, outputs or performance by exploiting the system vulnerabilities. [...]

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training dataset ('**data poisoning**'), or pre-trained components used in training ('**model poisoning**'), inputs designed to cause the model to make a mistake ('**adversarial examples' or 'model evasion**'), **confidentiality attacks** or model flaws.

Annex IV(2g) – Technical documentation

(g) the validation and testing procedures used, including information about the **validation and testing data** used and their main characteristics; **metrics** used to measure accuracy, robustness and compliance with other relevant requirements set out in Title III, Chapter 2 as well as potentially discriminatory impacts; test logs and all test reports dated and signed by the responsible persons, including with regard to pre-determined changes as referred to under point (f).

Article 52c(1) – Obligations for providers of general purpose AI models

(a) draw up and keep up-to-date the technical documentation of the model, including **its training and testing process and the results of its evaluation**, which shall contain, at a minimum, the elements set out in Annex XX for the purpose of **providing it, upon request, to the AI Office** and the national competent authorities;

(b) draw up, keep up-to-date and make available information and documentation to providers of AI systems who **intend to integrate the general-purpose AI model in their AI system**. Without prejudice to the need to respect and protect intellectual property rights and confidential business information or trade secrets in accordance with Union and national law, the information and documentation shall:

(i) enable providers of AI systems to have a good understanding of the **capabilities and limitations** of the general purpose AI model and to comply with their obligations pursuant to this Regulation; and [...]

↪ shall **not apply** to providers of AI models that are made accessible to the public under a **free and open licence** that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exception shall not apply to general purpose AI models with systemic risks.

(c) put in place a policy to **respect Union copyright law** in particular to identify and respect, including through state of the art technologies, the reservations of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;

(d) draw up and make publicly available a **sufficiently detailed summary about the content used for training** of the general-purpose AI model, according to a template provided by the AI Office;

Article 52d(1) – Obligations for providers of general-purpose AI models with systemic risk

(a) **perform model evaluation** in accordance with **standardised protocols and tools reflecting the state of the art**, including conducting and documenting **adversarial testing** of the model with a view to identify and mitigate systemic risk;

[...]

(d) ensure an **adequate level of cybersecurity protection** for the general purpose AI model with systemic risk and the physical infrastructure of the model.

Annex IXa Section 1(1) – Technical documentation (GPAI models)

1. A general description of the general-purpose AI model including:

a) **the tasks that the model is intended to perform** and the type and nature of AI systems in which it can be integrated;

b) acceptable use policies applicable;

c) the date of release and methods of distribution;

d) **the architecture and number of parameters**;

e) modality (e.g. text, image) and format of inputs and outputs;

f) the license;

Annex IXa Section 1(2) – Technical documentation (GPAI models)

2. A detailed description of the elements of the model refered to in paragraph 1, and relevant information of the process for the development, including the following elements:

a) the technical means (e.g. instructions of use, infrastructure, tools) required for the general-purpose AI model to be integrated in AI systems;

b) the design specifications of the model and training process, including **training methodologies and techniques**, the key design choices including the rationale and assumptions made; **what the model is designed to optimise for** and the relevance of the different parameters, as applicable;

c) information on **the data used for training, testing and validation**, where applicable, including type and provenance of data and **curation methodologies** (e.g. cleaning, filtering etc), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and **methods to detect identifiable biases**, where applicable;

d) the **computational resources used to train the model** (e.g. number of floating point operations – FLOPs-), training time, and other relevant details related to the training;

e) known or estimated **energy consumption** of the model; in case not known, this could be based on information about computational resources used;

Annex IXa Section 2 – Technical documentation (GPAI models with systemic risk)

3. Detailed description of **the evaluation strategies, including evaluation results**, on the basis of **available public evaluation protocols and tools** or otherwise of other evaluation methodologies. Evaluation strategies shall include evaluation criteria, metrics and the methodology on the identification of limitations.

4. Where applicable, detailed description of the measures put in place for the purpose of conducting **internal and/or external adversarial testing** (e.g., red teaming), **model adaptations, including alignment and fine-tuning**.

Where applicable, detailed description of the system architecture explaining **how software components build or feed into each other** and integrate into the overall processing.

Practical contributions to AI standardization

ISO TC 37

“*Language and Terminology*”

ISO IEC JTC 1/SC 42

“*Artificial intelligence*”

TC 37

“*Language and Terminology*”

JWG 5

JTC 1/SC 42

“*Artificial intelligence*”

ISO TC 37

"*Language and Terminology*"

JWG 5

ISO IEC JTC 1/SC 42

"*Artificial intelligence*"

cen CENELEC JTC 21

"*Artificial intelligence*"

ISO TC 37

“Language and Terminology”

JWG 5

ISO IEC JTC 1/SC 42

“Artificial intelligence”

EUROPEAN LANGUAGE GRID

SacreBLEU

cen CENELEC JTC 21

“Artificial intelligence”

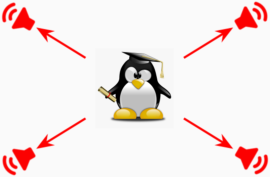
ISO TC 37

“Language and Terminology”

JWG 5

ISO IEC JTC 1/SC 42

“Artificial intelligence”

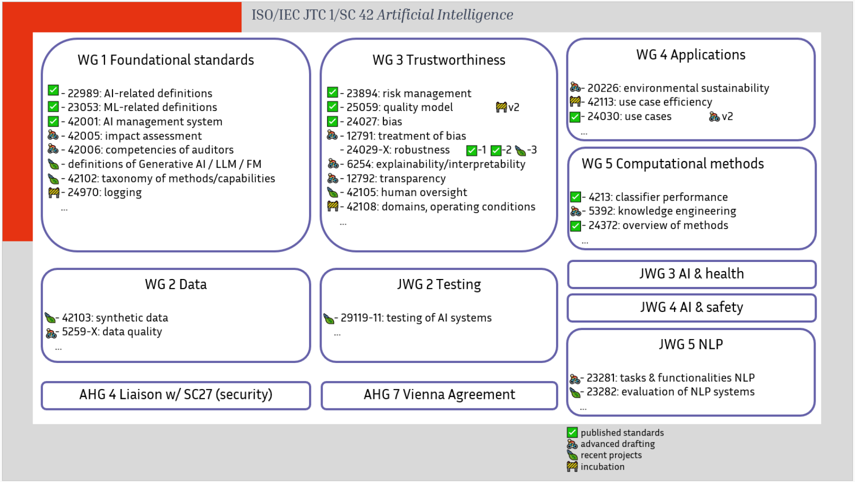


EUROPEAN LANGUAGE GRID

SacreBLEU

cen CENELEC JTC 21

“Artificial intelligence”



ISO/IEC JTC 1/SC 42 *Artificial Intelligence*

WG 1 Foundational standards

✅ - 22989: AI-related definitions
✅ - 23053: ML-related definitions
✅ - 42001: AI management system
🚴 - 42005: impact assessment
🚴 - 42006: competencies of auditors
🌱 - definitions of Generative AI / LLM / FM
🌱 - 42102: taxonomy of methods/capabilities
🚧 - 24970: logging
...

WG 3 Trustworthiness

✅ - 23894: risk management
✅ - 25059: quality model 🚧v2
✅ - 24027: bias
🚴 - 12791: treatment of bias
- 24029-X: robustness ✅-1 ✅-2 🌱-3
🚴 - 6254: explainability/interpretability
🚴 - 12792: transparency
🌱 - 42105: human oversight
🚧 - 42108: domains, operating conditions
...

WG 4 Applications

🚴 - 20226: environmental sustainability
🚧 - 42113: use case efficiency
✅ - 24030: use cases 🚴v2
...

WG 5 Computational methods

✅ - 4213: classifier performance
🚴 - 5392: knowledge engineering
✅ - 24372: overview of methods
...

WG 2 Data

🌱 - 42103: synthetic data
🚴 - 5259-X: data quality
...

JWG 2 Testing

🌱 - 29119-11: testing of AI systems
...

JWG 3 AI & health

JWG 4 AI & safety

JWG 5 NLP

🚴 - 23281: tasks & functionalities NLP
🌱 - 23282: evaluation of NLP systems
...

AHG 4 Liaison w/ SC27 (security)

AHG 7 Vienna Agreement

✅ published standards
🚴 advanced drafting
🌱 recent projects
🚧 incubation

CEN-CENELEC JTC 21 *Artificial Intelligence*

WG 1 Strategic Advisory Group

✅ - work programme AI Act & architecture
🌱 - technical coherence forum
🚴 - inclusiveness
🚴 - horizontal/vertical interplay
...

WG 2 Operational aspects

🚴 - 17894: conformity assessment
🚧 - AI-RM: risk management
...

WG 3 Engineering aspects

🚴 - 12791: bias
🚧 - 24970: logging
🚴 - 23281: tasks & functionalities NLP
🌱 - 23282: evaluation of NLP systems
🚧 - requirements for bias
🚧 - datasets
🚧 - computer vision
...

WG 4 Foundational & societal

🌱 - AI Trustworthiness framework
= entry point AI Act (harmonized standard)
🚴 - AI-enhanced nudging
🚧 - AI ethicists
🚴 - green AI
🚧 - fundamental rights impact assessment
...

WG 5 AI Cybersecurity

🚧 - complementing ISO standards
🚧 - AI-specific technical solutions

✅ published standards
🚴 advanced drafting
🌱 recent projects
🚧 incubation

Questions?

Thanks to The Noun Project for the wonderful icons!