Natural Language Processing, Standardisation, and the EU AI Act

James Davenport

University of Bath
ISO-IEC JTC1 SC42 JWG5 etc.
CEN-CENELEC JTC21 WG3 Convenor
Partially supported by STANDICT
All views expressed are personal

20 September 2025

REGULATION (EU) 2024/1689
OF THE EUROPEAN PARLIAMENT
AND OF THE COUNCIL
of 13 June 2024
(but drafting preceded the ChatGPT era)

- 44 pages of 180 recitals
- 74 pages of 101 articles
- 5 pages of supplementary articles
- 20 pages of Annexes

Generally enters into force 2 August 2026 (but prohibited uses on 2 February 2025, and some other special cases).
Many more systems/uses of systems fall/may fall into the "high risk" category (Annexe III) than one might believe.

Key Definitions

provider means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;

deployer means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity;

These pervade the Act, but aren't in general use outside this framework. A provider might also be a deployer.

~~NLP, Standardisation, and~~ the EU AI Act: Article 13 Transparency and provision of information to deployers

1. High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured . . .
2. High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers.
3. The instructions for use shall contain at least the following:
   (b) the characteristics, capabilities and limitations of performance of the high-risk AI system, including:
   (ii) the level of **accuracy**, including its metrics, robustness and cybersecurity[. . .]against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;

~~NLP, Standardisation, and~~ the EU AI Act: "accuracy"

JHD Consider a cancer which occurs in 1%. Two tests exist.

A. Always no cancer.
B. "Possible cancer" if there, but also on 2% of cases where not present.

A is more accurate (1% error rate, versus B's 1.98% error rate) but completely useless.

EC (§2.6 of Annex to standardisation request)

"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.

"Standardisation" is a phrase that many use, but comparatively few understand. Most developed, and many other, countries have standardisation bodies (generally one, Germany has two).

USA	ANSI = American National Standards Institute.
UK	BSI = British Standards Institute
France	AFNOR = Agence Français pour la NORmalisation
Germany	DIN and VDE. (BSI is cybersecurity agency)

In general, these are independent bodies, though they can receive "requests" from their national government.
These are members (full members for developed countries, but various "associate" status are possible) of international standardisation bodies.

International Standards Bodies (relevant)

IEC The International Electrotechnical Commission held its inaugural meeting on 26 June 1906

ISO The International Organization for Standardization was founded on 23 February 1947

Is computing an electrotechnical subject or not? Lengthy discussions, culminating in

JTC 1 Joint Technical Committee 1, entitled "Information technology", which was created in 1987

SC42 Sub-Committee 42, entitled "Artificial Intelligence", was created in 2017

JWG5 Joint Working Group ISO/IEC JTC1/SC42 - ISO/TC 37 WG: Natural language processing

European Standards Bodies (relevant)

CEN European Committee for Standardization (French: Comité Européen de Normalisation) was founded in 1961. 34 member countries (including UK)

CENELEC European Committee for Electrotechnical Standardization (French: Comité Européen de Normalisation ÉLECtrotechnique) was founded in 1973 (as a merger)

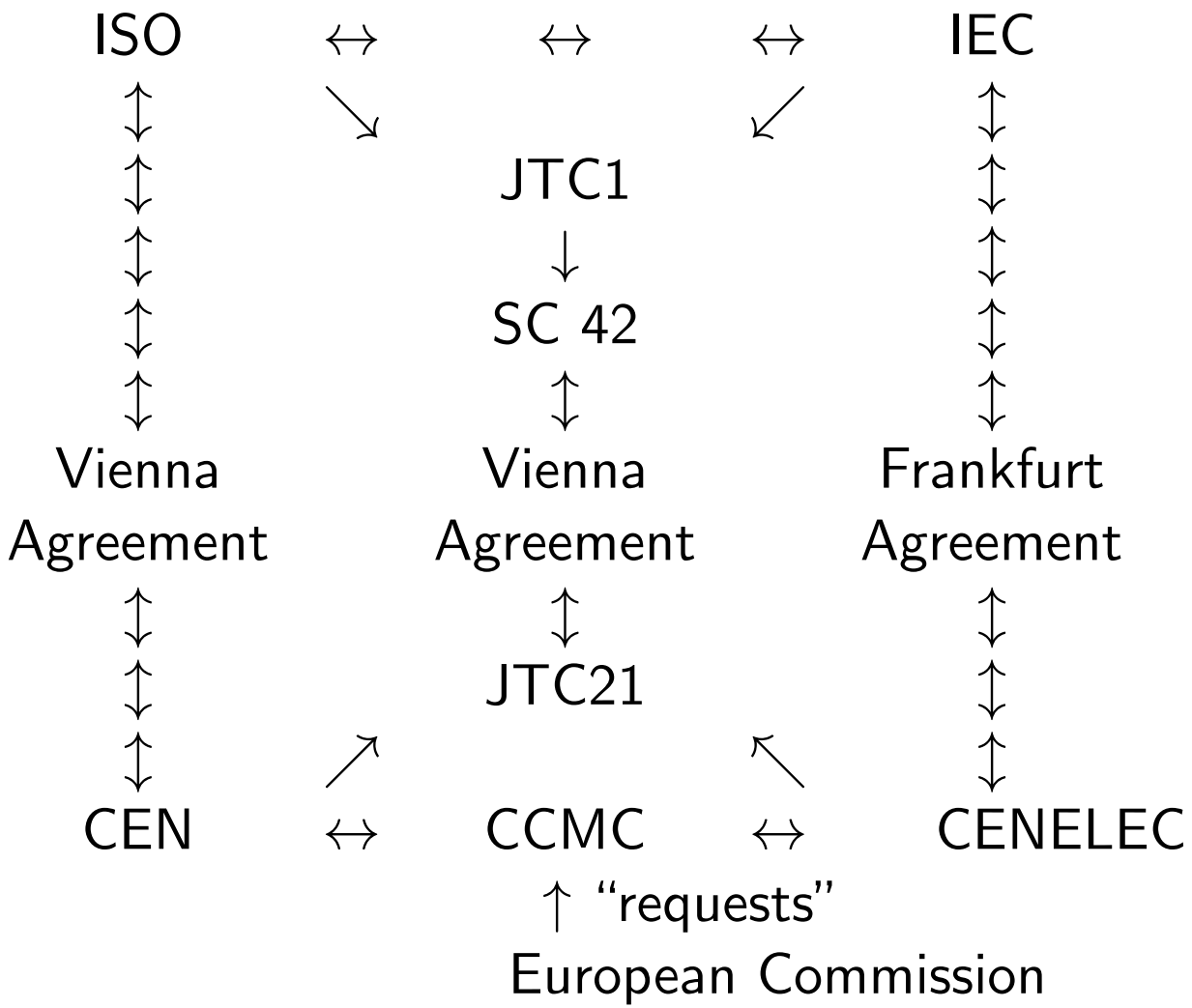
CCMC CEN–CENELEC Management Centre

ETSI European Telecommunications Standards Institute was set up in 1988 by the European Commission, and has a very different constitution.

Is computing an electrotechnical subject or not? Lengthy discussions, culminating in many Joint Technical Committees, including:

JTC 21 Joint Technical Committee 21, entitled "Artificial Intelligence", which was created by CEN & CENELEC in 2021.

Structure


ISO
IEC
JTC1
SC 42
Vienna Agreement
Vienna Agreement
Frankfurt Agreement
JTC21
CEN
CCMC
CENELEC
"requests"
European Commission

~~NLP~~, Standardisation, ~~and the EU AI Act~~: Example

Car seat belts (a policy issue, but technical details): UK example

- 1955± Discussions in BSI about seat belts.
- 1960 BS 3254 Specification for Seat Belt Assemblies for Motor Vehicles.
- 1966 Compulsory in front seats of all new cars (certified with designated approval mark, i.e. BSI)
- 1983 Front seat wearing compulsory.
- 1987 Compulsory in all seats of all new cars
- 1991 Compulsory wearing in all seats; 3254:1991 replaces 3254:1988 which replaced . . .
- 2002 BSI 3254 replaced by UN/ECE Regulation 16

Standards Publications/ Levels

There are various types of international/european standards (and many national ones!)

TR Technical report. No requirements. Often definitions, descriptions of issues etc.

TR 22989:2023 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.

TS "A Technical Specification addresses work still under technical development, or where it is believed that there will be a future, but not immediate, possibility of agreement on an International Standard. " [ISO/IEC, not CEN]

IS International Standard. Can contain requirements: "shall" etc.

EN European Norm. Equivalent of IS. Many IS are "adopted" or "adapted" as ENs.

hEN "harmonised EN", published on OJEU with Annex ZA explaining which clauses of the standard give a presumption of conformity with which pieces of European legislation.

~~NLP, Standardisation, and the~~ EU AI Act: 13.3(b)(ii) "The level of accuracy"

There will be a (multi-part) standard addressing Article 13, intended to be a hEN.
The "accuracy" clause will repeat §2.6: "accuracy" shall be understood as referring to the capability of the AI system to perform the task for which it has been designed.
So we need "tasks" as well as "metrics".

Many ISO-IEC 4213 — Performance measurement for AI classification, regression, clustering and recommendation tasks.

Vision Image recognition etc. Standards being developed in JTC21/WG3: a TR "taxonomy of tasks" and an EN for metrics of accuracy.

NLP Standards in JWG5: a TR 23281 "taxonomy of tasks" and an IS 23282 for metrics of accuracy.

Ideally, a standard has a nice simple requirement, and a corresponding method of testing it.
"A seat belt clip shall not open under a force of NNN Newtons, distributed in any way across the two straps. This will be tested by NNN Newtons on each strap, and by an equally distributed test".
But there is the "teaching to the test" problem, which is why human natural language examinations have "unseen translations" etc. See also "Dieselgate", where cars were, essentially, programmed to recognise the test track.

NLP, Standardisation, and the EU AI Act: 23281

Tasks related to shallow analysis of natural language contents

Tasks related to author identification and profiling

Tasks related to document analysis and mining

Tasks related to corpus-level analysis and mining

Tasks related to semantics and meaning

Tasks related with user interactions

Tasks involving the generation of linguistic contents

Tasks involving conversions among modalities

NLP, Standardisation, and the EU AI Act: 23281 Tasks related to shallow analysis of natural language

Sample task: language identification (from list). Modality specific

Text language identification : Given plain text written in a single language, decide whether that language is in the list, and if so which one it is.

Spoken language identification : Given one or more utterances spoken in the same language, decide whether/ which.

Sign language identification : Given a video recording of sign language content, decide whether/ which.

Document image language identification : Given an image of a document with content in a single language, decide whether/ which.

Scene text detection and language identification : Given an image or video in which text appears, produce the bounding box of each text region (+ timestamps, if video), and label each one with the language it is written in.

NLP, Standardisation, and the EU AI Act: Tasks related to shallow analysis of natural language (2)

Note that the provider chooses the list of languages.
This means that the deployer has to decide "what is a language".
Vlaamse/Nederlands for a comparatively uncontentious example: there are others in Europe, even close to here, which are more contentious.
What about American/British English?

US Davenport: a type of sofa that can be converted into a bed

UK Davenport: a type of portable writing desk

Note also the explicit "none of these": in general AI has not been developed to be good at saying "don't know":
`https://arxiv.org/abs/2509.04664`.

NLP, Standardisation, and the EU AI Act
23282: Artificial Intelligence — Evaluation methods for accurate natural language processing systems

Still very much "work in progress": there was intended to be a consultation of National Standards Bodies starting last week, but it has been delayed by work on 23281.
How are we going to evaluate?

Automated methods

Human evaluation protocols

Multi-dimensional analysis criteria "This includes quality criteria, but also error criteria"

NLP, Standardisation, and the EU AI Act: 23282

- Tasks related to shallow analysis of natural language contents
- Tasks related to author identification and profiling
- Tasks related to document analysis and mining
- Tasks related to corpus-level analysis and mining
- Tasks related to semantics and meaning
- Tasks related with user interactions
- Tasks involving the generation of linguistic contents
- Tasks involving conversions among modalities

NLP, Standardisation, and the EU AI Act: 23282 Tasks related to shallow analysis of natural language

Sample task: language identification (from list).

\+ This is a classification task, so falls under ISO 4213, and the measures defind there: Accuracy, FP, FN, F_1 etc.

? But the real question is about the test.

Naïvely, if I have k languages, I should test M_k times on samples of N_k words.

! And these samples should be genuine human-produced sentences, and by training/testing rules, independent of training data.

– Not feasible: I can tell English/French with five words, but American/British English is much harder.

?? This is currently an unknown area (to JWG5: any ideas/contributions welcomed)

NLP, Standardisation, and the EU AI Act: BLEU

6.2.7.1.2 When applying an n-gram or subsequence matching metric, the following information shall be specified

a) whether multiple reference sequences are used per candidate sequence, and how many.

6.2.7.1.3 When applying an n-gram or subsequence matching metric to text sequences (i.e. sequences of tokens), the following information shall be specified

a) whether the matching and the aggregation of n-gram counts are case-sensitive or case-insensitive;

b) the tokenization applied to the candidate and reference sequences, which shall be the same;

c) any lemmatization or stemming processing applied to the candidate and reference sequences before computation, including reference to the specific method or model used;

NLP, Standardisation, and the EU AI Act: BLEU (2)

6.2.7.1.3 continued

d) whether rare words are mapped to a special "unknown" token before computation;

e) whether specific elements are ignored in the sequences (e.g. punctuation, stop words).

6.2.7.2.4 When applying BLEU, the information in 6.2.7.1.2/3 and the following information shall be specified

a) the maximum n-gram length n. A common choice is 4;

b) when applied at the level of a dataset, whether the aggregated metric is computed as a macro-average, a macro-average with corpus-level brevity penalty or a micro-average.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

NLP, Standardisation, and the EU AI Act: 23282 Tasks involving the generation of linguistic contents: Machine translation

Various metrics are described.

BLEU can be applied to evaluate machine translation,

Calibration BLEU scores lower than 10 are usually not perceived as translations.
BLEU scores higher than 40 are usually perceived as high-quality translations.
Typical values for the news genre in well-resourced languages are between 30 and 45 BLEU.
BLEU scores for target languages with rich morphology are usually lower, for the same perceived quality.

Also a description of various human evaluation protocols.