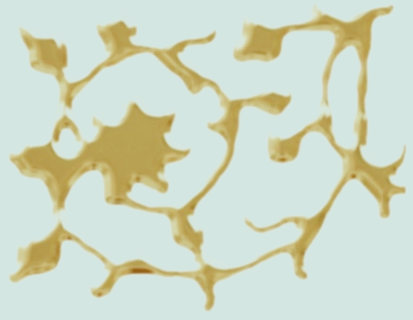




L'IA Act nomme les biais, pas la discrimination linguistique

L'IA Act ne crée pas de mécanisme explicite et spécifique de prévention de la discrimination linguistique. Cet angle mort du cadre européen de l'IA discrimine.

Christelle Dauvergne, Jurilangue

Juriste-linguiste • ES/CAT › FR



ARTICLE 5 DE L'IA ACT

Le critère de la langue considéré puis écarté

Considérant 16

Définit la catégorisation biométrique. Énumère les catégories possibles :

sexe, âge, tatouages, **langue**, religion, **appartenance à une minorité nationale**, orientation sexuelle ou politique...

→ **La langue est identifiée comme critère de catégorisation.**

**Art. 5(1)(g)**

Interdit la catégorisation biométrique fondée sur :

race, opinions politiques, affiliation syndicale, convictions religieuses ou philosophiques, vie sexuelle, orientation sexuelle.

→ **La langue et l'appartenance à une minorité ont disparu.**

Le législateur identifie la langue mais choisit de ne pas la protéger.



LE DÉCALAGE NORMATIF

Deux cadres, deux logiques

Charte des droits fondamentaux de l'UE

Article 21

Interdit la discrimination fondée sur la langue.

Article 22

Impose le respect de la diversité linguistique.

≠

IA Act

(Règlement 2024/1689)

Aucun mécanisme explicite et spécifique ne prévoit l'évaluation de la discrimination linguistique dans les systèmes d'IA.

La langue n'est pas un critère de biais à surveiller dans les données d'entraînement (art. 10).

Le préambule de l'IA Act affirme que le règlement doit être mis en œuvre conformément à la Charte. Pourtant, il ne traduit pas ses articles 21 et 22 en obligations opérationnelles.



LES 7 PRINCIPES ÉTHIQUES DU HLEG*

Un principe prometteur sans traduction concrète

1. Action humaine et contrôle humain
2. Robustesse technique et sécurité
3. Vie privée et gouvernance des données
4. Transparence

5. Diversité, non-discrimination et équité

← ANGLE MORT

6. Bien-être sociétal et environnemental
7. Responsabilité

Ce principe devrait couvrir la diversité linguistique. Or, en l'absence de critères de couverture linguistique dans les actes délégués et les normes harmonisées, il n'est suivi d'aucun effet.

*Groupe d'experts de haut niveau de la Commission européenne



CONVENTION-CADRE DU CONSEIL DE L'EUROPE**

Le pont juridique existe

STCE n° 225 – Ratifiée par l'UE le 15 mai 2026

Article 10**Égalité et non-discrimination**

Renvoie au droit international applicable. L'article 14 de la CEDH interdit notamment les discriminations fondées sur la langue dans la jouissance des droits garantis par la Convention.

Article 17**Non-discrimination**

« Sans discrimination fondée sur quelque motif que ce soit » – formulation plus large que l'IA Act.

Article 16**Gestion des risques**

Exige la consultation des « parties prenantes pertinentes, en particulier les personnes dont les droits pourraient être affectés ».

Convention-cadre → CEDH (art. 14) → discrimination linguistique dans les systèmes d'IA : le pont juridique est tracé. Il manque sa traduction opérationnelle.

** Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit



LE BIAIS SYSTÉMIQUE

Un déséquilibre que les modèles ne corrigent pas

Part de l'anglais dans les corpus d'entraînement

GPT-3 (OpenAI, 2023)

92,65 %

Les autres langues représentées dans les données se partagent les 7,35 % restants.

Llama 2 (Meta, Touvron et al., 2023)

89,70 %

Ex. : catalan : 0,04 % (16^e position sur 27 langues recensées). Les 27 langues recensées sont celles dépassant 0,005 %.

Corpus OSCAR (web scraping)

Anglais : 328 Md de tokens

Catalan : 3,5 Md (×94 moins) • Basque : 0,3 Md (×1 000 moins)

Et les autres ?

Anthropic (Claude) et Mistral ne publient aucune distribution linguistique de leurs corpus. Les langues minoritaires ne figurent même pas dans leurs benchmarks.

« Our training data is not evenly distributed across languages. Some languages have far more data than others. »

Anthropic, Claude's values across models and languages, 13 juillet 2026

L'IA Act exige la transparence des modèles à usage général (art. 53). Or, sans critère de couverture linguistique, cette transparence est vaine. On ne saurait corriger un biais qu'on ne mesure pas.

Une langue qui disparaît des données disparaît du monde

La diversité linguistique est un trésor de l’humanité. Elle porte nos identités, nos histoires, nos droits et nos manières de comprendre le monde.

Les corpus d’entraînement sous-représentent massivement certaines langues, avec des écarts pouvant atteindre un rapport de un à mille. Ce déséquilibre reste insuffisamment mesuré, rarement comparable et souvent opaque. Certains fournisseurs ne publient pas la composition linguistique de leurs données.

Une langue n’est pas un simple canal d’information : elle est une manière de découper le réel, de nommer les liens de parenté, de dire le droit et de porter une mémoire. Lorsqu’une langue est marginalisée dans les données, c’est une manière de penser et de transmettre le monde qui devient moins visible et moins accessible dans les outils numériques. Ce patrimoine appartient à toutes et tous, d’où que l’on vienne et quelle que soit la langue que l’on parle.

L’Europe s’est dotée du premier cadre juridique horizontal et contraignant au monde consacré à l’intelligence artificielle. Il lui reste à protéger, dans les systèmes d’IA, ce que l’Europe a de plus ancien : la pluralité de ses langues.

Christelle Dauvergne

Juriste-linguiste • Jurilangue • jurilangue.fr

© 2026 Christelle Dauvergne – Jurilangue. Reproduction autorisée avec mention de la source.

