Why Translation Errors Must Be Part of the AI Safety Discussion

Labeling, use-case typologies, and risk mitigation for AI translation

TRANQUALITY

Why Translation Errors Must Be Part of the AI Safety Discussion

Labeling, use-case typologies, and risk mitigation for AI translation

Authors

Alan Melby, PhD

Ryan Foley

Publisher

Tranquality, a division of LTAC Global

tranquality.info | LTAC Global

Date

September 2025

Suggested citation

Melby, A., & Foley, R. (2025). *Why Translation Errors Must Be Part of the AI Safety Discussion: Labeling, use-case typologies, and risk mitigation for AI translation*. Tranquality (LTAC Global), September 2025.

About Tranquality

Tranquality is a division of LTAC Global focused on research, standards, and practical guidance for safe, high-quality language solutions in high-impact domains.

Acknowledgments

The authors thank colleagues in standards bodies, research groups, and language-access coalitions whose incident reporting and analyses informed this report.

Disclaimer

This publication is for informational purposes only and does not constitute legal advice. Organizations should consult counsel regarding regulatory obligations in their jurisdictions.

Contact

info@tranquality.info

Table of Contents

Executive Summary

AI translation and interpreting now mediate trillions of cross-language interactions each day, but mainstream AI safety frameworks seldom treat translation failure as a first-order risk. This report argues that translation is a high-impact AI system in many workflows and that its distinctive failure mode—correspondence errors, where fluent output misrepresents source meaning—creates a hidden risk that monolingual users cannot detect.

We synthesize evidence of scale, documented incidents, as well as equity impacts, and we propose a practical, label-first governance model that pairs a use-case typology and gating rules with two verification statuses from ASTM International: **Professionally Verified Translation** and **Unverified Translation**. The aim is simple: make provenance visible before anyone acts, route high-impact decisions to qualified bilingual review, and prevent silent propagation of errors at scale.

Key findings of this report

- **Translation errors as an AI-safety issue.** In high-impact contexts, a single mistranslation can precipitate clinical, legal, financial, operational, or safety harm.
- **Correspondence errors as the dominant source of consequential harm.** Output that is fluent yet wrong in meaning, misrepresenting the source text, is undetectable to monolingual users and is amplified by fluency bias.
- **Systemic risk at scale.** Even low error rates yield large numbers of consequential mistakes when billions of words are auto-translated daily.
- **Disproportionate impact on Indigenous and under-resourced languages.** AI translation quality significantly drops off for over 7000 languages outside a small set of 17 “Tier-One” languages, increasing risk to users from many language groups.
- **Gate high- vs low-impact scenarios.** A use-case typology and gating rules clarify when **Professionally Verified Translation** is required and when **Unverified Translation** may be used for triage or drafting with guardrails.
- **Labels as a first line of defense.** A dual-label regime, stacked with AI-Generated Content (AIGC) disclosure, restores informed consent, enables auditability, and supports procurement and platform policy.
- **Actionable path forward.** Puts translation errors on the AI-safety agenda, requires clear disclosure of raw AI translations, and labels every translation.

1. Translation and AI Safety

AI safety is incomplete until translation errors are part of the conversation.

Safety discourse around artificial intelligence now occupies center stage in academic, business, and policy circles, yet the conversation remains curiously silent on AI-generated translation, a technology whose early iterations predate, and in many ways enabled, today's large-language models (LLMs). Long before transformer architecture could generate photorealistic images or draft legal briefs, neural machine translation (NMT) systems were quietly translating trillions of words per day, shaping everything from social-media feeds to e-commerce.

1.1 Missing in Paris AI safety report: translation

Despite the ubiquity of AI translation, AI safety discussions have been silent on its dangers in high-risk scenarios. The recent International AI Safety Report 2025 (Department for Science, Innovation and Technology & AI Safety Institute, 2025), hereafter referred to as the "Paris Report," describes scenarios arising from malicious actors, malfunctions, bias, loss of control, and systemic risks emerging from widespread deployment of general-purpose AI models. Examples are provided for each category; however, the report makes no mention of the risks arising from AI translation errors. That omission matters because unpredictable translation errors are inherent in the technology and have already produced serious harmful outcomes (Appendix A). AI translation and its emerging sibling, AI interpreting, are not low-stakes conveniences, or solved problems, but are high-impact AI systems (Simard, 2024) – systems with modes of failure that are not fully covered by standard AI-safety taxonomies.

The Paris Report organizes risk reasoning around a Capabilities → Risks → Risk-Management pipeline. Applying this frame to language services reveals an immediate blind spot. AI translation capabilities have advanced from rule-based substitutions to contextual, transformer-driven systems that routinely outperform bilingual novices. Yet, these same systems introduce a blind spot: *correspondence errors*, fluent outputs that misrepresent intended source meaning, and thus remain invisible to monolingual users. Classical AI-safety taxonomies recognize risks posed by "hallucinations" and "bias" but not with regard to translation, leaving regulators unprepared for translation-specific harms.

1.2 Why the blind spot persists

The marginalization of translation within AI-safety discussions stems from three misconceptions:

- **Perceived maturity.** Because MT has been commercially deployed for decades, and made huge progress, policymakers assume that AI translation is a solved problem, even though high-stakes errors persist (Pym, 2025).
- **Fluency.** Modern MT and other AI translation output often *looks* perfect, giving a false impression of correspondence to the source (Pym, 2025). However, correspondence cannot be verified or fact-checked by an end-user who cannot understand the language of the source text.
- **Invisible Automation.** Browsers, apps, and platforms often auto-translate without clear notice, undermining informed consent (Simard, 2024).

2. Evidence of High Risk

AI safety frameworks support the claim that translation errors are high-risk in specific use cases.

International regulators increasingly rely on structured tests to decide whether an AI system merits the "high-impact" or "high-risk" designation that triggers enhanced governance. Canada's proposed *Artificial Intelligence & Data Act* (AIDA) articulates seven risk-weighted factors (Innovation, Science and Economic Development Canada, 2025);[1] the EU AI Act and the OECD's risk framework apply similar criteria. When AI translation systems are evaluated against these frameworks, they qualify unambiguously as high-risk systems when used in high-risk scenarios, sitting squarely in the same regulatory class as autonomous vehicles and diagnostic decision support systems. Safety-critical translation errors, liability gaps, and data-security exposure are well-documented in AI translation practice (Canfora & Ottmann, 2020).

Box 2-1. AIDA's seven factors applied to AI translation

1. **Evidence of risk –** Documented mistranslations in asylum, emergency, and medical settings already threaten life, liberty, and due-process rights.
2. **Severity of harm –** A single erroneous phrase in healthcare, law enforcement, or supply chain communication can precipitate wrongful surgery, detention, or loss of life.
3. **Scale of use –** Billions of daily AI-translated interactions replicate any systemic defect worldwide in seconds.
4. **Observed incidents –** Vacated verdicts, diplomatic tensions, and irreversible patient outcomes stem from unvetted AI translation in official channels.
5. **No opt-out –** Users encounter AI translations (often undisclosed) in social media, e-commerce, and public services, making avoidance impossible.
6. **Vulnerable groups –** Speakers of under-resourced or Indigenous languages face higher error rates and fewer review options, widening existing inequities.
7. **Regulatory gap –** Current medical device processes, consumer-protection regulation, and AI policies do not address linguistic accuracy or transparency, leaving systemic risks from AI translation and interpreting systems largely unchecked.

The following analysis illustrates how, in many cases, AI translation and interpreting systems would meet the threshold for "high-impact" designation in emerging legislation such as Canada's proposed AIDA.

[1] Government of Canada, *Artificial Intelligence and Data Act* (Bill C-27, 2024 draft).

2.1 Scale

AIDA Factor 3 – Scale of use

Automatic translation and interpreting are now woven into everyday digital infrastructure, from social media timelines to cross-border supply chain dashboards. Recent estimates place the aggregate daily volume of raw machine-translated text above 1 trillion words, a figure that dwarfs the human capacity for post-editing or verification, a high-risk factor called *volume beyond human oversight*. In practical terms, this means that even a low error rate of one substantive mistranslation per 10,000 sentences can yield tens of thousands of high-stakes errors every 24 hours. These errors may be further propagated in use cases where unverified translation output is distributed at scale digitally.

2.2 Documented harms

AIDA Factors 2 and 4 – Severity of harm; **Nature of harm already observed**

Case law and incident reports document several recurrent harm vectors leading to adverse clinical events, incorrect judicial rulings, misleading emergency messaging alerts, compromised operational safety instructions, and reputational crises predicated on mistranslated source material (Canfora & Ottmann, 2020). Collectively, these harms are *foreseeable* and *systemic*, satisfying AIDA's requirement that a high-impact system imposes "non-negligible risk of serious harm."

Malicious misuse of AI translation systems, including espionage and the deliberate insertion of false translations, *does* exist, but current evidence shows that unintentional malfunctions already occur at scale and generate most of the observed harm, and are therefore the focus of this report.

2.3 Equity and unequal burden

AIDA Factors 1 and 3 **– Evidence of risk; Disproportionate impact on vulnerable or disadvantaged groups**

Quality gains accrue disproportionately to a small set of 17 "Tier-One" languages, including English, Chinese, and Japanese, that attract the bulk of training data and investment. Public CSA Research materials place Tier One at roughly 90% of global GDP among internet users (Lommel, 2024).

For the world's other 7,000-plus languages, automated translation quality typically drops, as shown in Translation Quality Evaluation (TQE), by lower scores on standard MT evaluation benchmarks and human assessments. These gaps reflect data scarcity and

model performance limits documented in large-scale multilingual research (Robinson et al., 2023).

Communities using under-resourced or Indigenous languages also face fewer options for qualified human review of AI translation, which compounds error rates and restricts meaningful access to essential services. In the US, inadequate language access is linked to longer hospital stays and higher revisits/readmissions for people with limited English proficiency. Federal rules under Section 1557 of the Affordable Care Act require covered entities to take reasonable steps to ensure meaningful access and to avoid discrimination when deploying AI-enabled tools.

In Europe, both the European Convention on Human Rights (ECHR) (Article 14) and the EU Charter (Article 21) prohibit discrimination on grounds including *language*. The EU Charter (Article 35) recognizes the right of access to preventive healthcare and medical treatment. Separately, ECHR Article 6(3)(e) guarantees the right to an interpreter in criminal proceedings. Together, these frameworks signal that language barriers can function as rights-denying conditions across critical services, reinforcing the need to treat language as a protected asset in AI impact assessments and governance.

2.4 Asymmetry of correspondence errors vs monolingual errors

***AIDA Factors 1,2, and 4* – Evidence of risk; *Severity of harm;* Nature of harm already observed**

Conventional quality metrics reward fluency, yet correspondence errors, semantically plausible target language output that diverges materially from the source, dominate consequential harm from AI translation systems.

Two attributes intensify risk due to correspondence errors:

1. **Asymmetric bilingual scenario**: Users of translated content are normally monolingual and thus unable to assess quality. Even when users have access to the source text, they typically do not speak or read that language well enough to be able to verify correspondence and fact-check AI-generated output as they might in question-answer, summarization, or other monolingual applications of GenAI.
2. **Fluency bias**: NMT and GenAI systems produce output that reads naturally, lulling readers into accepting incorrect content. Fluent output can mask errors and raise detection costs, amplifying risk in high-stakes use cases (Pym, 2025)

Correspondence errors exploit exactly the form of hidden-failure mode that warrants heightened governance.

2.5 Labels as a first line of defense

AIDA Factor 5 – Inability to opt out

Web browsers, social platforms, and mobile apps often auto-translate without explicit notice in what Simard (2024) defines as *Invisible automation*. This undermines informed consent, since end-users may act on unverified content under the assumption that it is original or professionally vetted. Labeling unvetted AI translations can help mitigate this risk.

ASTM International is in the process of approving a draft amendment to standard F2575 that adopts a dual-label regime as a risk mitigation strategy:

- **Professionally Verified Translation**: bilingual expert review completed; liability traceable.
- **Unverified Translation**: no qualified review; includes raw AI translation and non-qualified human output.

Labeling neither cures bias nor corrects correspondence errors, but it *inserts an explicit decision point* into otherwise invisible workflows, potentially breaking a silent propagation chain that can scale translation errors.

Global policy signal. The most current and comprehensive approach to labeling AI-generated content was launched by China in September 2025, setting a global precedent for visible and embedded markings on AI-created media. This initiative mandates that all forms of AI content, including text, images, video, and audio, must include both explicit, visible labels and implicit digital watermarks or cryptographic metadata in the file itself. (Digital Watch Observatory, 2025)

Policies like these carve out a path for AI content labeling and make stacking both an AI-Generated Content (AIGC) disclosure with ASTM translation-provenance labels practical.

We expand on labeling as a risk mitigation strategy in section four of this report.

3. Use-Case Typologies and Risk of Harm

Not all translation use cases carry the same risk. What matters is the consequence of acting on a potential error and whether professional translation verification is available before any consequential step. Table 3-A groups common scenarios accordingly: **high-impact** when a mistranslation can affect health, liberty, enforceable obligations, or public safety; **low-impact/easily mitigated** when the only action is triage for probable relevance or when verification is readily available and required prior to action. Box 3-1 summarizes the gating rules that determine which path applies.

This typology can be used to determine when the use of raw output of AI-generated or other unverified translation is acceptable versus when professionally verified translation is essential. Table 3-A also indicates when to incorporate escalation paths, and when "human-at-the-core" checkpoints should be incorporated into the translation workflow to minimize the risk of decision-makers acting on erroneous information.

Box 3-1. Gating rules for unverified translation

Use **Professionally Verified Translation** vs **Unverified Translation** according to these gates:

1. **Consequence gate.** If a decision can affect health, liberty, enforceable obligations, or safety, require Professionally Verified Translation before any action.
2. **Reuse and exposure gate.** If output will enter a system of record, be sent outside the team, or be reused beyond triage or drafting, require Professionally Verified Translation.
3. **Verification gate.** Unverified Translation is acceptable only for triage, low-visibility content, or internal drafting when a qualified bilingual reviewer is required and available to verify before any consequential step. Quarantine and label Unverified Translation until verified.
4. **Language and domain gate.** Treat low-resource languages, specialist terminology, or known low Translation Quality Evaluation (TQE) contexts as high impact; require Professionally Verified Translation.
5. **Audit and labeling gate.** Always label Unverified Translation and log routing to Professionally Verified Translation for traceability.

Table 3-A. Use-case typology by decision impact and required gate (see Box 3-1)

Adapted from risk-management strategies and automation tradeoffs in Pym (2025).

High-impact use cases	Low-impact/easily mitigated use cases
Clinical care and life safety - diagnosis, treatment plans, medication instructions, discharge summaries, emergency alerts, environmental/industrial safety notices. Downstream actions can directly affect health or physical safety. *Professionally Verified Translation required before action*.	**Triage for probable relevance** - the *only* decision is whether content *might* be relevant and therefore merits professional translation (e.g., inbox pre-sorting, document discovery, open-source intelligence pre-screening).
Due process and liberty - asylum interviews, police reports, custody recommendations, plea agreements, parole/probation instructions, immigration detention materials, prison safety guidance. *Professionally Verified Translation required before action*.	**No decision required** - informational or entertainment uses where users are not asked to act (e.g., subtitles for entertainment, casual browsing of user-generated content).
Binding legal or financial effects - contracts, procurement specs, insurance policies, consent forms, product warranties, audit findings, compliance attestations, trade or other legal compliance documentation. *Professionally Verified Translation required before action*.	**Ready verification before action** - a qualified bilingual reviewer is readily available and *required* to verify correspondence *before* any consequential step (e.g., internal drafts routed to *Professionally Verified Translation* gating; customer-support responses parked until verified).
Critical infrastructure and operations - aviation/transport notices to airmen/mariners, utility switching procedures, industrial SOPs, hazard signage, labeling for drugs and medical devices. *Professionally Verified Translation required before action*.	**Ephemeral internal awareness (no operational effect)** - internal chatter or FYI updates where unverified text is quarantined from operational systems and cannot trigger binding actions (with auto-deletion or escalation to *Professionally Verified Translation* gating if reused).
Education and public services with rights attached - special-education plans, school disciplinary notices, benefits eligibility letters, health-plan communications, voting instructions. *Professionally Verified Translation required before action*.	**Search, retrieval, and summarization with guardrails** - features explicitly flagged as AI-translated, used only to *locate* materials, with any substantive use routed to *Professionally Verified Translation* gating.
Vulnerable populations or low-resource languages - when TQE is known to be low or specialist terminology is high (rare diseases, legal jargon), or where power asymmetries prevent remedy after error. *Professionally Verified Translation required before action*.	**Multilingual UX elements** - navigational scaffolding *that does not alter terms of service, rights, or safety* (e.g., menu labels), with rollback and professional review paths.

3.1 Domains of observed harm

While risk is present across domains, some high-stakes domains experience elevated risk of safety, legal, and operational harm.

Healthcare. Unverified translations have altered instructions, obscured contraindications, and contributed to consent failures. In clinical discharge materials, small shifts (omitted negation, dropped dosage qualifiers, or swapped technical terms) can escalate to readmissions or irreversible outcomes. Many of these errors are *correspondence errors*.

Justice, immigration, and corrections. Misinterpretations of pronouns, place names, or regulatory acronyms have jeopardized asylum claims, distorted legal histories, and undermined detainee protections. When AI-mediated interpreting omits salient background cues, officers and victims face avoidable risk.

Public warnings & operational safety. Pym documents a bushfire instruction sequence in which the clause "If you are caught in fire in your car," followed by "position the car facing the fire," was repeatedly mistranslated by major MT/LLM systems as "If your car is on fire," rendering the subsequent action nonsensical and unsafe.

In the same set, "Try to position your car towards the approaching fire" surfaced in Chinese as "尝试将汽车停在靠近火场的位置" (back-translation: "Try to park your car close to a fire"), an inversion that can turn a public alert into a hazard at population scale (Pym 2025).

Illustrative incidents aligned to the domains above are summarized in Table 3-B, with case notes in Appendix A.

***Table 3-B*. Selected incidents by domain (case notes in Appendix A)**

Domain	Incident	Consequence
Healthcare	Unverified medical translations altered care instructions.	Medication non-adherence; patient safety risk.
Justice	AI-assisted translation tablets used for legal aid services missed critical details during client consultations.	Erroneous legal advice; Due-process risk.
Immigration	Asylum claim jeopardized by pronoun/place-name errors.	Deportation risk; rights deprivation.
Public warning	Critical public safety instructions during wildfire were mistranslated in emergency messaging.	Community safety risk.
Operational safety	Safety label mistranslated in device documentation.	Recalls and operational hazards.

4. Label-First Mitigation

The goal of using standard labels is to make provenance visible *before* anyone can act on a translation. A dual-label regime of **Professionally Verified Translation** and **Unverified Translation** is being codified by ASTM International, paired with an "AI-Generated Content" disclosure creates an auditable decision point that interrupts the silent propagation of correspondence errors and restores informed consent. Labels do not fix bias or semantic drift by themselves; they reveal provenance, assign responsibility, and gate consequential use. Labeling translations by provenance helps users calibrate trust before action (Pym, 2025).

4.1 What the labels mean and why they matter

Professionally Verified Translation.
Indicates that a *qualified* bilingual professional has verified the translation against agreed project specifications, and that accountability is traceable to a named verifier/reviewer or provider. Professionally Verified Translation is appropriate for high-impact scenarios identified in Section 3.

Unverified Translation.
Indicates that no qualified bilingual verification has been performed. Unverified Translation covers (a) raw AI/MT output and (b) output produced by non-qualified translators. Within the framework of this report, Unverified Translation is permissible only in low-impact/easily mitigated use cases (see Section 3), *and only when guardrails are in place*. Guardrails include explicit disclosure, quarantine from operational systems, and ready paths to escalate to Professionally Verified Translation before any consequential action is taken.

Why this works.
Web/app auto-translation often occurs without explicit notice; users then act on unverified content under the assumption it is original or professionally vetted. Visible labels insert a decision point and restore informed consent.

Default presumption for users. In the absence of a Professionally Verified Translation label, users can assume a translation is raw/unverified.

Practically, this aligns with real-world usage patterns whereby the overwhelming majority of translations users encounter are raw or otherwise unverified. Decision-makers should therefore treat an unlabeled translation as an Unverified Translation.

4.2 Chain of trust: process, qualifications, accountability

Process. Labels are process-agnostic: they apply to fully human translation, post-edited MT, and GenAI-prompted translation. What matters is the presence of *verification status* at document delivery.

Qualifications. "Qualified" means competence in the working languages and the domain, and the ability to verify compliance with specifications: accuracy, completeness, terminology, style, and any regulatory constraints. A Professionally Verified Translation mark attaches to the document only when a professional has performed and documented verification (confirmation that the translation meets agreed-on specifications, regardless of the process used to obtain it).

Accountability. Labels originate with the professional translator, reviser, or editor who verifies the work and are *passed forward* by the publisher (the organization making content available to end users), who bears ultimate responsibility for what consumers read. This chain of custody is the practical safeguard that converts labels from a UI badge into enforceable governance.

Implication. Because raw AI output and non-qualified human output are indistinguishable to most consumers, both are labeled Unverified Translation to alert users.

5. Recommendations

The evidence is clear: AI translation and interpreting systems operate inside high-impact workflows, and their distinctive failure modes, especially correspondence errors, are underrepresented in mainstream AI safety guidance. We recommend three near-term actions that regulators, publishers, and providers can implement now:

1) Put translation errors on the AI safety agenda.

- Add cross-lingual failure modes, including correspondence errors, to AI risk frameworks, red-team protocols, and incident taxonomies.
- When a system touches health, due process, binding obligations, public safety, or fundamental rights, treat AI translation as high-impact scenario by default; use the use-case typology in Section 3 of this report to gate decisions and require escalation to qualified review.
- Stand up incident capture and disclosure channels so cases are not siloed and can inform continuous improvement.

2) Mandate labeling of raw automatic translation using existing AI-generated content mechanisms.

- Require a clear "AI-Translated" disclosure wherever raw MT or GenAI output is shown to end users, leveraging the same technical and policy infrastructure used for AIGC.
- Pair the disclosure with a use warning in high-impact contexts (e.g., "Do not rely on translated output without professional verification") and provide an easy path for users to request Professionally Verified Translation in workflows.
- Preserve both AIGC and translation verification status labels across platforms and APIs; do not allow UI themes, copy-paste, or export pipelines to strip labels.

3) Label every translation as Professionally Verified Translation or Unverified Translation.

- Adopt ASTM labels across all channels:
    - **Professionally Verified Translation** when a qualified bilingual professional has verified the translation against agreed specifications and accountability is traceable.

    - **Unverified Translation** when no such verification has occurred, including raw AI output and non-qualified human output.

- Make labels visible to people and machines; preserve them through publication and downstream processing, and audit label changes in production logs.

Implementation note. Procurement and platform policies should make Professionally Verified Translation mandatory in the high-impact scenarios and allow Unverified Translation only for low-impact scenarios with guardrails. These steps provide consumer protection, clarify accountability, and insert a decision checkpoint where invisible automation currently operates.

References

ASTM International. (2024, October 10). WK92487: Revision of F2575-23e2 Standard Practice for Language Translation [Work item]. ASTM International. Retrieved September 15, 2025, from https://www.astm.org/membership-participation/technical-committees/workitems/workitem-wk92487

BBC News. (2023, November 21). *NHS interpreting service problems contributed to patient deaths*. https://www.bbc.com/news/uk-england-bristol-66605536

Bhuiyan, J. (2023, September 7). *Lost in AI translation: Growing reliance on language apps jeopardizes some asylum applications*. The Guardian. https://www.theguardian.com/us-news/2023/sep/07/asylum-seekers-ai-translation-apps

Canfora, C., & Ottmann, A. (2018, July). *Of ostriches, pyramids, and Swiss cheese: Risks in safety-critical translations* [Preprint]. ResearchGate. https://www.researchgate.net/publication/326317704

Canfora, C., & Ottmann, A. (2020). Risks in neural machine translation. *Translation Spaces, 9*(1), 58–77. https://doi.org/10.1075/ts.00021.can

Council of Europe. (latest consolidated). *European Convention on Human Rights.* https://www.echr.coe.int/documents/d/echr/convention_ENG

Deck, A. (2023, April 19). *AI translation is jeopardizing Afghan asylum claims*. Rest of World. https://restofworld.org/2023/ai-translation-errors-afghan-refugees-asylum/

Department for Science, Innovation and Technology (DSIT) & AI Safety Institute. (2025, January 29). *International AI Safety Report 2025* (DSIT Research Paper Series 2025/001). GOV.UK. https://www.gov.uk/government/publications/international-ai-safety-report-2025/international-ai-safety-report-2025

Digital Watch Observatory. (2025, September 2). AI-generated media must now carry labels in China. https://dig.watch/updates/ai-generated-media-must-now-carry-labels-in-china

Eby, H., Green, C., & Hylak, B. (2024, August 20). *Human vs. machine interpreting today: A professional assessment* [Report]. Interpretation Professional Advisory Committee (IPAC). https://www.ata-divisions.org/ID/wp-content/uploads/2024/09/20240906_machine_interpreting_today_v2.pdf

European Union. (2012). *Charter of Fundamental Rights of the European Union (2012/C 326/02). Official Journal of the European Union*, C 326, 391–407. https://eur-lex.europa.eu/eli/treaty/char_2012/oj

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union*, L 2024/1689 (12 July 2024). ELI: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Innovation, Science and Economic Development Canada. (2025, January 31). *The Artificial Intelligence and Data Act (AIDA) – Companion document.* Government of Canada. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

Justice Innovation. (n.d.). *Tag: Cristina Llop*. Retrieved September 15, 2025, from https://justiceinnovation.law.stanford.edu/tag/cristina-llop/

Khoong, E. C., Steinbrook, E., Brown, C., & Fernandez, A. (2019). *Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. JAMA Internal Medicine, 179*(4), 580–582. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2725080

Lester, G. (2025, February). *Translation errors in ICE detention settings* Lester, G. (2025, February). *Translation errors in ICE detention settings* [Conference presentation]. Interagency Language Roundtable (ILR) Virtual.

Lommel, A. (2024, November). How does wealth affect language access? CSA Research. Retrieved September 15, 2025, from https://insights.csa-research.com/reportaction/305013650/Toc

OECD (2024), *Framework on management of emerging critical risks* (*OECD Public Governance Policy Papers*, No. 49). OECD Publishing, Paris, https://doi.org/10.1787/2f2eddd8-en

O'Shea, J. (2021). How legal documents translated outside institutions affect lives, businesses and the economy. *International Journal for the Semiotics of Law*, 34, 1331-1373.

Pielmeier, H. (2024, September). *Automated speech-to-speech interpreting: Six evaluation dimensions for professional deployments*. CSA Research. https://insights.csa-research.com/reportaction/305013706/Toc

Pym, A. (2025). *Risk Management in Translation*. Cambridge University Press. https://doi.org/10.1017/9781009546836

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023, December 6–7). *ChatGPT MT: Competitive for high- (but not low-) resource languages*. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 392–418). Association for Computational Linguistics. https://www2.statmt.org/wmt23/pdf/2023.wmt-1.40.pdf

Simard, M. (2024). Position paper: *Should machine translation be labelled as AI-generated content?* In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (AMTA 2024): Volume 1, Research Papers* (Chicago, IL, Sept 30–Oct 2, 2024). National Research Council Canada.

Stakeholders Advocating for Fair and Ethical AI in Interpreting (SAFE AI) Task Force, & Coalition for Sign Language Equity in Technology (CoSET). (2025, August). AI interpreting solutions evaluation toolkit: Part A: Organization, implementation and management. https://safeaitf.org/wp-content/uploads/2025/09/AI-Interpreting-Solutions-Evaluation-Toolkit_Part-A.pdf

Stakeholders Advocating for Fair and Ethical AI in Interpreting (SAFE AI) Task Force. (2024, June). *Interpreting SAFE AI Task Force Guidance: AI and interpreting services*. SAFE AI Task Force. https://safeaitf.org/wp-content/uploads/2024/07/SAFE-AI-Guidance-07-01-24.pdf

The Japan News. (2025, February 13). *Senkaku Islands Referred to as Diaoyu Islands in Subtitles of NHK Report; Japanese Broadcaster Says AI System Responsible*. The Japan News by The Yomiuri Shimbun. https://japannews.yomiuri.co.jp/society/general-news/20250213-238395/

United Nations, High-level Advisory Body on Artificial Intelligence. (2024, September). *Governing AI for humanity: Final report***. United Nations.** https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

U.S. Department of Health and Human Services. (2024, May 6). *Nondiscrimination in Health Programs and Activities (Final Rule). Federal Register*, 89(88), 37522–37854. https://www.govinfo.gov/content/pkg/FR-2024-05-06/pdf/2024-08711.pdf

Appendices

Appendix A. High-Consequence Translation Errors

Year	Domain	Consequence	Error/Failure
2022	**Acute healthcare**	Patient consent incorrectly obtained via Google Translate; uterus removed without fully comprehensible briefing, raising informed-consent concerns (BBC, 2023). *Note: Phone interpreters engaged after 15 minutes; quality issues continued.*	**High-stakes risk** / accuracy errors and interactional miscommunication
2020	**Asylum adjudication**	Claim rejected after MT swapped first-person singular for plural pronouns ("I" → "we"), creating inconsistencies between oral and written testimony (Rest of World, 2023).	**Correspondence error** / pronoun ambiguity
2019	**Asylum adjudication**	Claim jeopardized, MT mistranslated city name ("Belo Horizonte" → "beautiful horizon"), distorting personal history (Bhuiyan 2023).	**Factual error** / proper noun mistranslation
2020	**Broadcast journalism**	NHK (Japan Broadcasting Corporation) discontinued its AI-based multilingual subtitles service after its Google Translate-powered system mistakenly rendered 'Senkaku Islands' as 'Diaoyu Islands,' the Chinese designation, during a live English broadcast. The error raised diplomatic and accuracy concerns, prompting the broadcaster to end the service (The Japan News, 2025).	**Factual error** / politically sensitive proper-noun (toponym) mistranslation
2019	**Clinical discharge instructions**	Unedited MT translations contained life-threatening errors in Spanish (2%) and Chinese (8%) versions of emergency department discharge instructions, directly jeopardizing patient safety (Khoong et al., 2019).	**High stakes risk** / unedited medical translation errors
2021	**Commercial contracts**	MT expanded a Greek corporate abbreviation into an unrelated foreign entity, potentially invalidating a contractual agreement (O'Shea, 2021).	**Factual error** / entity mistranslation
2025	**Correctional guidance**	Prisoner protection guidelines obscured; Prison Rape Elimination Act acronym PREA rendered as *preá* ("guinea pig") in Brazilian Portuguese; detainees filed formal grievances (Lester, 2025).	**Factual error** / acronym mistranslation
2023	**Emergency messaging**	MT error changed critical instruction from "position your car towards the approaching fire" to "park your car close to a fire," potentially life-threatening misguidance (Pym, 2025 – p.48).	**Factual error** / hazardous instruction mistranslation
2023	**Emergency messaging**	MT mistranslated conditional safety instruction ("if you are caught in fire in your car") as "if your car is on fire," causing instructions for action to become	**Correspondence error** / conditional misinterpretation

		nonsensical and endangering lives (Pym, 2025 p.43-44).	
2023	**Emergency messaging**	MT randomly alternated between formal and informal second-person registers, diminishing trustworthiness and clarity in urgent public safety instructions (Pym, 2025 p. 29-44).	**Pragmatic error** / **Correspondence error** / register inconsistency
2025	**Emergency response (911) / Law enforcement**	Deputies arrived unaware of a volatile situation on the scene of a domestic violence case due to AI-interpreted 911 call omitting critical background context (sounds indicating escalating dispute); victim sustained head injury, case currently under litigation. (A. Birchfield, President and Legislative Chair, AAPTI, personal communication, March 12, 2025).	**Correspondence error** / Omission of critical contextual cues
2019	**Healthcare instructions**	MT omission altered meaning in English-to-Dutch translation, removing the advice to inform the doctor about existing stomach or intestine issues, increasing risk to patient safety (Daems & Macken, 2019, as cited in Canfora & Ottmann, 2020 – p. 61).	**Correspondence error** / critical detail omission
2019	**Immigration detention**	Six-month communication blackout for detainee due to voice-translation tool misrecognizing Afro-Indigenous Portuguese accent; medical issues went untreated (Bhuiyan 2023).	**High-stakes risk** / Speech-to-Text accent bias
2024	**Immigration detention**	Detainee visitation rights jeopardized; Spanish translation of Immigration and Customs Enforcement handbook (ICE) drops negation ("cannot" → "can"), telling visitors that they ***can*** wear controversial and obscene prints, (Lester, 2025).	**Correspondence error** / omission of negation
2025	**Legal aid services**	AI-assisted translation missed critical details during client consultations ("I never received the notice"), resulting in erroneous legal advice and potential infringement of due process (Justice Innovation, n.d.).	**Correspondence error** / critical detail omission

Appendix B – Glossary of terms

AI translation & interpreting (MT, NMT, GenAI, MI): Machine-mediated cross-language transfer for text or speech. MT = machine translation; NMT = neural machine translation; GenAI = generative AI; MI = machine interpreting.

Correspondence error: Output that is fluent yet wrong in meaning, misrepresenting the source text or speech.

Fluency bias: The tendency to trust natural-sounding output even when the underlying meaning is incorrect.

Auto-translation (invisible): Automatic translation applied without clear user notice or labeling, creating a risk that readers act on unverified content.

High-impact scenario: A use case in which acting on a translation error can affect health, liberty, enforceable obligations, or public safety.

Use-case typology and decision gating: The report's framework for determining when **Professionally Verified Translation** is mandatory and when **Unverified Translation** may be used for triage or drafting with guardrails.

Labels (Professionally Verified Translation and Unverified Translation): Verification-status marks that indicate whether a qualified professional has reviewed the translation before use. Pair these with a visible AI-Translated disclosure whenever raw machine output is shown to end users.

Provenance and chain of trust: The metadata and audit trail that bind a translation to its source, model or engine, human verifier, timestamp, verification step, and applied labels to support accountability.

Tier-One languages: High-resource languages that benefit from abundant training data and investment, leading to generally better model performance.

TQE (Translation Quality Evaluation): Methods and metrics used to assess translation quality. Note that high aggregate scores can mask correspondence errors, so high-impact use cases still require qualified review.