ISSN 2616-7433 Vol. 7, Issue 7: 9-14, DOI: 10.25236/FSST.2025.070702

AI Translation and Data Ownership: Ethical Conflicts, Legal Risks, and Translator Agency in the Algorithmic Era

Chen Ziwei

Fuzhou Technology and Business University, Fuzhou, China

Abstract: *With the widespread integration of generative AI into translation practices, questions surrounding the legitimacy of data acquisition, translator marginalization, and ethical restructuring have become increasingly urgent. Anchored in the question "Whose data?", this paper investigates the mechanisms of data utilization in AI translation systems and the ethical conflicts they generate. It identifies key risks, including unauthorized data extraction, lack of informed consent, and structural linguistic inequality. The study further reveals that translators' intellectual labor is frequently de-personalized and stripped of attribution, leading to a loss of professional agency. These issues are particularly pronounced in legal translation, where terminological inaccuracies, ambiguous accountability, and hallucinated outputs pose heightened risks. In response, this paper proposes a governance framework centered on traceability, distributed accountability, and informed authorization. Ultimately, it argues for the role of AI as a tool that assists rather than replaces human translators, and calls for a collaborative, ethically grounded translation ecosystem.*

Keywords: *AI translation; data rights; translation ethics; translator marginalization; legal translation; governance framework*

1. Introduction

The rapid advancement of artificial intelligence is reshaping the landscape of the language service industry. In the domain of translation, large-scale pre-trained language models such as ChatGPT and DeepL have increasingly replaced human translators in a wide range of tasks. While AI translation technologies offer undeniable advantages in efficiency and accessibility, they also provoke urgent ethical concerns: Where does the training data come from? Is it legally authorized? Does it violate the rights of data subjects?[1-5]

The notion of translation ethics was initially conceptualized by French translation theorist Antoine Berman in a lecture series titled La Traduction et la Lettre at the Collège International de Philosophie[6]. Since then, the topic has garnered increasing scholarly interest across both Western and Chinese academia[7-10].

In their study on translation data ethics in the era of big data, Wang and Liu highlight two pressing issues: data alienation and data misuse. On one hand, translation data is stripped of its contextual and cultural features during standardization and computational processing. On the other hand, training AI models with such data—often without consent from translators or copyright holders—amounts to an infringement of data rights, posing serious ethical challenges to the profession[11].

To address these emerging dilemmas, Zhang and Qu introduces the notion of "translation technology ethics", emphasizing the need to evaluate translation practices through a tripartite lens of technology–human–society. He argues that translation ethics in the age of AI should encompass not only the interaction between human translators and machines but also the legitimacy, transparency, and control of data usage, as well as the broader socio-ethical implications of translation technologies[12].

Within this context, this study centers on the critical question of "Whose data?". It explores the ethical tensions surrounding data ownership and usage in AI translation, particularly focusing on the legality of training data sources, the marginalization of translators' intellectual and moral rights, and the potential reconfiguration of the translation ecosystem. Ultimately, the aim is to propose a framework for fair, accountable, and human-centered data governance in AI translation, contributing to the safeguarding of translators' agency and the integrity of professional ethics in the digital age.

ISSN 2616-7433 Vol. 7, Issue 7: 9-14, DOI: 10.25236/FSST.2025.070702

2. Mechanism of AI Translation and Data Source Analysis

2.1 Neural Network-based Translation Technologies

The rapid development of artificial intelligence has driven translation systems into a new phase based on deep learning. Mainstream AI translation systems such as ChatGPT, Google Translate, and Wenxin Yiyan are built upon Transformer-based large language models, trained on massive bilingual or multilingual corpora to enable automatic interlingual conversion.

These models employ self-supervised learning and reinforcement learning from human feedback (RLHF) to optimize translation outputs via contextual modeling and intent recognition[13-14]. However, their powerful performance relies heavily on the quality and diversity of training data, raising widespread concerns over data legitimacy, controllability, and ethics.

As Wang points out, although generative AI translation technologies have achieved breakthroughs in efficiency, their training data often suffer from ambiguity of origin, leading to ethical risks such as bias, alienation, and lack of authorization[15].

2.2 Classification of Data Sources for AI Translation Systems

These ethical concerns become more tangible when examining the various categories of training data employed by AI translation systems. As summarized in Table 1, different data types—ranging from public web content to user-generated translations—entail distinct ethical risks.

Table 1. Ethical Risks Associated with AI Translation Data Sources

Data Type	Examples	Ethical Risk
Public web content	Wikipedia, government websites, multilingual regulations	Unauthorized use, copyright ambiguity
Open-access corpora	OPUS, Europarl, UN Corpus	No explicit permission for usage
User-generated translations	Google Translate community, Facebook crowdsourcing	Lack of informed consent from users
Data mined from platforms	Translation software input/output logs	privacy concerns and original content appropriation

From an ethical perspective, even if the data is “technically available,” using it without informed consent or explicit licensing constitutes a violation of data rights and translator authorship[16].

Han further emphasizes that the healthy development of translation technology must follow three core principles of business ethics: legality, legitimacy, and transparency; otherwise, it may intensify the technological alienation of translation labor[10].

2.3 Legal Controversies and Case Studies

Although some platforms claim to use “public data,” many legal and ethical controversies remain. Typical examples include:

Case 1:

Training AI models using published translations (e.g., novels) without the consent of authors or translators violates copyright law.

Case 2:

User-generated translation data is logged and used for future training without informed consent.

Zhang and Qu argues that "the development of translation technologies must not come at the expense of translator rights; rather, a multi-stakeholder, legally responsible governance mechanism should be institutionalized"[12].

Furthermore, Wang and Liu criticize AI systems for over-relying on data from dominant languages, thereby marginalizing minority languages and cultures—an effect they term "digital linguistic colonialism"[11].

The preceding chapter examined the operational logic and linguistic datasets underpinning AI-

ISSN 2616-7433 Vol. 7, Issue 7: 9-14, DOI: 10.25236/FSST.2025.070702

driven translation systems. Beyond recognizing notable improvements in processing efficiency, it also brought attention to a range of unresolved issues concerning data authorization, the erosion of translator agency, and imbalances in linguistic representation. These considerations set the stage for a deeper inquiry into the ethical tensions explored in the following sections.

3. Ethical Conflicts in AI Translation – Data Rights and Translator Displacement

3.1 Ethical Ambiguity of Data Rights

Modern AI translation tools rely heavily on extensive multilingual data collections. This reliance brings forward a host of complex ethical questions about who possesses legal and moral authority over such content. Although translated texts have long been regarded as a form of intellectual property, in AI development contexts they are frequently extracted without consent and reappropriated for commercial deployment, thereby depriving translators of control and proper attribution over their original contributions.

Wang and Liu characterize this large-scale, unconsented use of textual data as a form of appropriation, constituting a dual infringement on the rights of both original authors and translators[11]. This ethical ambiguity is exacerbated by the opacity of data sources and the absence of clear legal or contractual frameworks governing the reuse of translation content in AI development.

3.2 Marginalization and "Invisibility" of Translators

Despite improvements in speed and scalability, AI translation technologies have diminished the role of human translators, reducing them from active interlingual mediators to passive post-editors—or, in some cases, rendering them entirely redundant. This not only undermines the translator's professional identity but also erases their presence from the chain of data value production.

Translators are subject to a dual form of ethical invisibility: the absence of attribution and the systemic **devaluation of their labor**. This invisibility is not simply a byproduct of technological substitution but stems from the broader process of datafication, which filters out the translator's creative agency, cultural fluency, and aesthetic judgment—elements that are essential yet unquantifiable in algorithmic terms.

3.3 Tensions Between Professional Ethics and Technological Use

The integration of generative AI into legal translation introduces complex ethical and professional dilemmas. Legal language is characterized by its precision, standardization, and binding implications. When AI systems—trained on generalized or unverified corpora—are used to translate legal documents, there is a heightened risk of semantic distortion, terminological inaccuracy, and factual "hallucinations," which may lead to severe legal consequences.

In contrast to general or literary translation, legal texts demand exact interpretative alignment and a nuanced understanding of contextual frameworks. Even minor inaccuracies can compromise the clarity of legal intent, violate contractual obligations, or trigger procedural injustice. Yet, AI systems do not possess the capacity for juridical inference or intercultural sensitivity, and they may generate linguistically coherent but substantively flawed output—thereby fostering a misleading perception of reliability.

The use of AI in such sensitive domains raises a fundamental ethical dilemma: Should speed and cost-efficiency take precedence over legal accountability and professional expertise? The indiscriminate deployment of AI in legal translation without rigorous human oversight undermines the ethical responsibility of both translators and the institutions that employ such tools. As Zhang and Qu argue, aligning AI applications with established ethical norms in specialized domains like law is not merely a technical task, but a moral imperative[12].

ISSN 2616-7433 Vol. 7, Issue 7: 9-14, DOI: 10.25236/FSST.2025.070702

4. AI Risks and Regulatory Proposals in Legal Translation

4.1 Terminological Misalignment and Semantic Ambiguity in Legal Translation

Legal language is defined by its precision, terminological specificity, and binding legal consequences. However, AI translation models frequently lack domain-specific knowledge and legal annotation capabilities, resulting in semantic ambiguity, lexical inaccuracies, and mistranslation of critical clauses, and these errors that may directly affect litigation outcomes or contractual validity[17].

There is an example that an AI system mistranslated the phrase “shall be liable for damages” as “shall bear damage”, failing to convey the legal distinction between liability and compensatory obligation. This misinterpretation could distort the enforceability of relevant legal provisions.

The key risks associated with AI-generated legal translations can be summarized as follows: First, misjudgment of terminological equivalence, where AI systems fail to capture nuanced legal distinctions between seemingly corresponding terms across languages. Second, over-simplification of legal clauses and syntactic structures, often stripping away legally significant nuances and complexities. Third, AI-generated “hallucinations”, wherein models produce legally non-existent or inaccurate content that appears plausible but lacks legal basis.

To address these challenges, AI applications in legal translation should integrate juridical validation engines capable of referencing authoritative legal lexicons and clause databases in real time. In parallel, it is crucial to include mandatory usage disclaimers clarifying that AI-generated outputs are strictly informational and lack any legally enforceable authority.

4.2 Traceability Issues and Accountability Gaps in Generative AI

Generative AI functions as a “black-box mechanism”, rendering it difficult to trace the origin of generated content. This lack of transparency complicates error attribution and conceals potentially illicit data usage during model training.

Wang observes that terminological biases and syntactic distortions in AI translations often stem from unverified or biased training data[15]. Nonetheless, platform providers rarely disclose corpus sources, leading to unresolvable accountability gaps.

For instance, in one documented case, an AI system translated the term "termination clause" as “termination explanation” , leading a client to overlook a critical clause necessary for enforcing breach of contract remedies. The platform involved declined responsibility, attributing the error solely to “automated algorithmic processing.”

To address these challenges, we should establish a mandatory training data registration mechanism to enhance corpus transparency and source verification; implement a risk-based labeling system for AI-generated outputs, introducing clear alerts for high-risk content such as legal translations; and develop a shared accountability framework that clearly defines the responsibilities of both platform providers and users of AI-generated translations.

4.3 Constructing Translator Data Rights and Consent Mechanisms

Translator-generated corpora are often utilized by AI platforms without prior consent, ignoring both the intellectual value and data subjectivity inherent in translators’ work. It is essential to institutionalize their rights to attribution, economic participation, and post-use revocation.

Zhang and Qu advocates for a data ethics framework that includes a **triadic mechanism**: informed consent prior to data use, proper attribution during use, and the right to withdraw data after deployment[12].

Legislative and Industry Recommendations:

a) Develop a sector-wide “Ethical Code for AI Translation Corpora Usage”

b) Require platforms to provide opt-in/opt-out mechanisms for corpus participation

c) Create a centralized “Translator Alliance Database” to record data usage access and enable revocation control

Although AI translation technologies significantly enhance efficiency and reduce costs, they also

ISSN 2616-7433 Vol. 7, Issue 7: 9-14, DOI: 10.25236/FSST.2025.070702

pose notable challenges, including potential legal misinterpretations and deeper ethical concerns embedded within algorithmic systems. Ensuring the responsible and sustainable application of AI in legal translation requires the implementation of comprehensive oversight mechanisms, transparent data management practices, and the protection of translators' legitimate rights.

Looking ahead, regulatory initiatives should strive for a careful balance between innovation and caution. The goal should be to construct an ethical framework for AI-mediated translation that is inclusive, anticipatory, and rooted in respect for human agency and the professional dignity of translators.

5. Conclusion and Future Outlook

5.1 Summary of Findings

This study, centered on the pivotal inquiry of "Whose data?", examined the ethical tensions surrounding data rights and translator marginalization in the context of AI translation. Through a detailed analysis of data sourcing and AI translation mechanisms, several key conclusions have been drawn:

Data alienation and rights erosion: The training data used in AI systems is often extracted, standardized, and applied without authorization, resulting in the systemic deprivation of translators' intellectual property and moral rights.

Degradation of professional ethics: AI technologies increasingly reduce translators to data suppliers or post-editors, thereby undermining professional recognition and diminishing the value of human linguistic labor.

Critical risks in legal translation: The application of AI in legal contexts entails heightened risks—ranging from semantic misrepresentation to ambiguous responsibility—which call for immediate regulatory intervention.

5.2 Limitations and Reflections

This research is primarily theoretical in nature and relies heavily on literature review; it lacks empirical data on specific AI translation platforms or measurable tracking of data usage pathways. Furthermore, the opacity of data pipelines in commercial AI models constrains deeper technical verification and limits case-specific analysis.

Future studies should incorporate empirical methods such as translator interviews, corpus traceability, and cross-platform comparative analysis. There is also a need to explore the differentiated impact of AI translation on professionals working across various languages, domains, and levels of expertise.

5.3 Future Directions

Policy Dimension: Promote the development of formal regulatory frameworks such as "Ethical Guidelines for Data Use in AI Language Technologies" and "Standards for AI Application in Legal Translation," to delineate the responsibilities of platforms, translators, and end-users.

Technological Dimension: Advance explainable AI (XAI) models with integrated traceability mechanisms, especially in sensitive fields such as legal and medical translation, to ensure transparency and mitigate algorithmic opacity[18].

Translator Empowerment: Explore the design of a "Translator Data Contract" system that guarantees translators' rights to consent, benefit financially, and revoke usage authorization over their translated content within AI ecosystems.

In an age where AI rapidly transforms the landscape of translation, data is no longer a neutral resource. It embodies power, carries ethical weight, and delineates the boundaries of governance. Translators, as agents of intercultural communication, must reclaim their voice, rights, and agency in the digital era.

The ultimate role of technology should be to assist, not replace; to empower, not alienate. Constructing a fair, transparent, and human-centered ethical paradigm for AI translation requires the

ISSN 2616-7433 Vol. 7, Issue 7: 9-14, DOI: 10.25236/FSST.2025.070702

joint participation of developers, translators, policymakers, and society at large.

References

[1] Cronin, M. Translation in the Digital Age[M]. London & New York: Routledge, 2013.
[2] Ensinger, B D, Presas, M. PACTE: Building a Translation Competence Model[A]. Alves, F. (ed). Triangulating Translation: Perspectives in Process-Oriented Research[C]. Amsterdam: John Benjamins, 2003: 134–141.
[3] Pym, A. What technology does to translating[J]. The International Journal for Translation and Interpreting Research, 2011(1): 1-9.
[4] Schäffner, C.Translation and norms[M].Beijing: Foreign Language Teaching and Research Press, 2007.
[5] Sela-Sheffy,Rakefet.How to be a (Recognized) Translator: Rethinking Habitus, Norms, and the Field of Translation[J].Target International Journal on Translation Studies, 2005, 17(1):1-26.
[6] Berman, A. La traduction et ses discours[J]. Meta, 1989, 34(4): 672-679.
[7] Bender E M, Gebru T, Mcmillan-Major A ,et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?[C].Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2021: 610–623.
[8] Floridi L, Taddeo M.What is data ethics?[J].Philosophical Transactions of the Royal Society A, 2016, 374(2083): 20160360.
[9] Kenny D, ed. Human Issues in Translation Technology[M]. London: Routledge, 2017.
[10] Han, L. Basic Principles of Commercial Ethics in Translation Technology from the Perspective of the Language Industry[J]. Shanghai Journal of Translators, 2019(5), 52–57.
[11] Wang, H. & Liu, S. Research on Translation Data Ethics in the Era of Big Data: Concepts, Issues and Suggestions[J]. Shanghai Journal of Translators, 2022(2), 12–17.
[12] Zhang, F. & Qu, X. Ethical Review of Legal Translation Technology[J]. Foreign Languages in China, 2021,18(6), 17–22.
[13] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022(35): 27730-27744.
[14] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 7871–7880.
[15] Wang, H. Ethical Risks of Generative AI Translation Technology in Legal Translation[J]. Journal of Gansu University of Political Science and Law, 2025(1), 1–10.
[16] Bowker L. Translation technology and ethics[M].The Routledge handbook of translation and ethics. Routledge, 2020: 262-278.
[17] Yu, X., Zheng, G., & Ding, X. Six Legal Issues of Generative Artificial Intelligence: A Case Study of ChatGPT[J]. China Law Review,2023 (2), 1–10.
[18] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[EB/OL]. arXiv:1702.08608, 2017[2025-08-17]. Available: https://arxiv.org/abs/1702.08608.