AMERICAN Journal of Science on Integration and Human Development
Volume 3, Issue 9, 2025 ISSN (E): 2993-2750

AI-Assisted Risk Assessment in National Security Translations

Manal ELtayeb Mohamed Idris
Department of Foreign Languages, Al-Baha University, Al-Baha, Kingdom of Saudi Arabia,
manalidris123m@gmail.com

Abstract: This paper examines the novel use of artificial intelligence (AI) in risk assessment frameworks for national security applications. Contemporary scholarship and practice indicate a significant deficiency: although AI-driven neural machine translation (NMT) is extensively utilized in U.S. intelligence and national security agencies (e.g., for Open-Source Intelligence [OSINT] and Signals Intelligence [SIGINT]), a standardized risk-aware methodology to assess the security, reliability, and error-propagation risks associated with these translations is conspicuously absent. This research presents a unique Risk-Aware Translation Framework (RATF) that integrates Probabilistic Risk Assessment (PRA) methodologies with AI-enhanced translation systems. The framework incorporates human-in-the-loop auditing and AI governance systems, providing a systematic enhancement to the discipline. Empirical evidence indicates that RATF markedly enhances the identification of semantic distortions, biases, and misclassification hazards. The ramifications are significant for U.S. national security entities, indicating the potential for AI-assisted translation to be utilized judiciously, with stringent control, to improve trust, precision, and robustness in multilingual intelligence endeavors.

Keywords: AI, Framework, Models, National security, Machine Translation, translation U.S, National security

Introduction

Due to the rapid proliferation of multilingual information on digital platforms, U.S. national security agencies are increasingly dependent on AI-assisted translation systems to analyze substantial volumes of foreign-language data. (Ibrahim, 2017). The U.S. intelligence community utilizes machine translation for OSINT surveillance, including social media, encrypted conversations, and diplomatic documents. However, while such systems are invaluable for speed and scale, they pose critical risks: mistranslations, semantic drift, and inadvertent disclosure of sensitive information.(Ibrahim, 2017) Recent research (Ibrahim, 2025; Hamdan, 2024; Brill, 2025) indicates that although AI-driven translation improves productivity, it does not possess a structured risk assessment framework capable of identifying and addressing possible risks. U.S. policy initiatives, like the National Artificial Intelligence Research and Development Strategic Plan (2023) and the AI Executive Order (Ibrahim, 2025; White House, 2023), emphasize the necessity for responsible AI use in national security; yet, translation risk assessment is still inadequately conceived. There is presently no systematic framework that merges probabilistic risk models with AI-enhanced national security translation. This paper presents the RATF framework, which modifies PRA principles for language technology. RATF offers a systematic, verifiable framework for risk detection, reduction, and governance within translation workflows. The findings will guide U.S. defense agencies, lawmakers, and translation experts, providing avenues for safer AI inclusion in multilingual security environments.

1.2 Statement of the Study Problem

Despite the growing integration of AI-driven translation in U.S. intelligence and defense activities, a systematic methodology for evaluating translation-related risks (such as semantic inaccuracies, security vulnerabilities, and hostile manipulation) remains absent. This absence engenders weaknesses in intelligence precision and policy execution.

1.3 The Aims of Study

The research seeks to establish the Risk-Aware Translation Framework (RATF) by modifying Probabilistic Risk Assessment (PRA) for AI-assisted translation. It aims to experimentally assess the framework in simulated U.S. national security translation scenarios and to investigate the impact of human-in-the-loop auditing on the reliability of AI-driven translations. The report seeks to offer policy recommendations for the incorporation of RATF into U.S. national security translation systems.

1.4 Questions of the Study

1. What is the approach for integrating PRA methodologies with AI-driven translation to improve risk detection in U.S. national security scenarios?
2. To what degree does RATF enhance the detection of semantic and contextual risks in comparison to traditional translation workflows?
3. What function does human-in-the-loop auditing serve in enhancing trust, precision, and accountability?
4. What governance frameworks are essential for the efficient implementation of RATF within U.S. intelligence agencies?

2-Literature Review

2.1 Probabilistic Risk Assessment (PRA) in AI Safet

Probabilistic Risk Assessment (PRA) has been utilized in high-reliability sectors such as nuclear power and aerospace, where failure implications are catastrophic and uncertainties require meticulous quantification. Recently, researchers have started to modify PRA approaches for the field of artificial intelligence, especially large language models (LLMs), where the risks are systemic, intricate, and frequently obscure. Wisakanto et al. (2025) present a PRA framework specifically designed for AI that combines hazard pathway modeling with uncertainty propagation, facilitating a more systematic assessment of vulnerabilities in general-purpose AI systems. He et al. (2025) presents probabilistic safety limitations for embodied AI, emphasizing the potential of PRA to extend beyond conventional engineering applications into dynamic, data-driven contexts.

PRA has certain benefits within the realm of national security. It enables policymakers to assess both the probability of AI system failures and the magnitude of their subsequent effects on information security and operational decision-making. (Ibrahim, 2022) Huang et al. (2024) assert that probabilistic approaches are a crucial adjunct to rule-based safety frameworks, facilitating the measurement of residual risk that remains post-implementation of safeguards. A recent survey by ScienceDirect (2025) illustrates the application of PRA models to LLMs for assessing safety, privacy, and adversarial vulnerabilities, which are increasingly pertinent in translation workflows involving classified or sensitive documents.

Furthermore, PRA's alignment with security frameworks like the CIA+TA paradigm (Confidentiality, Integrity, Availability + Trust and Autonomy) establishes it as a methodological conduit between AI governance and operational security (Aydin, 2025). The PRA-for-AI project (2025) offers toolkits that model failure pathways in translation-enabled AI systems, demonstrating how probabilistic modeling can predict mistranslation of essential intelligence documents or the breach of secret information. This methodology aligns with the extensive literature on security-by-design in large language models (Thompson, 2025), emphasizing that risk quantification should be conducted during the design process instead of being applied post-deployment.

This collection of work emphasizes that PRA is not simply an engineering-derived risk management tool, but a transformative technique for AI safety research. Its innovation resides in facilitating risk assessment within profound uncertainty, when both the model and its surroundings undergo unpredictable evolution (Ibrahim, 2022; Zhou et al., 2024). The ramifications for national security translation are significant: PRA can assist institutions in determining which AI-assisted translation processes meet established standards of secrecy and reliability, and which provide intolerable levels of systemic risk. (Ibrahim, 2019)

2.2 CIA+TA Framework in AI Governance

The CIA+TA paradigm signifies an advancement of the conventional cybersecurity triad—Confidentiality, Integrity, and Availability—by integrating two supplementary elements vital for the regulation of artificial intelligence systems: Trust and Autonomy (Crawford, 2025). This expansion recognizes that AI systems, especially large language models (LLMs), not only handle sensitive material but also function in semi-autonomous environments where human supervision may be restricted. Academics contend that this approach offers a more thorough basis for assessing AI governance, as it encompasses both technological vulnerabilities and socio-ethical hazards (Ibrahim, 2022; Huang et al., 2024).

Confidentiality is paramount in translation-enabled AI systems, especially in national security scenarios where mishandling confidential material may lead to significant breaches. Recent empirical research indicates that LLMs may inadvertently disclose training data, underscoring the necessity for robust confidentiality measures (Ibrahim, 2023; ScienceDirect, 2025). Integrity, denoting the trustworthiness and authenticity of outputs, is similarly crucial; misinterpretations or altered results can skew intelligence evaluations, as emphasized by WSJ (2025) in its inquiry of LLM abuse inside defense sectors. Availability guarantees that AI translation systems are resilient and impervious to denial-of-service or poisoning attempts, a risk increasingly recognized in extensive implementations (OWASP, 2025).

The incorporation of trust and autonomy differentiates the CIA+TA framework from preceding approaches. Trust underscores the necessity for transparency and responsibility in AI results. Aydin (2025) asserts that integrating trust measures into governance frameworks enables organizations to assess not just the operational efficacy of an AI system but also the trustworthiness of its judgments for human operators. Autonomy pertains to the extent to which AI systems operate independently in critical contexts. Holste and Agnetta (2025) contend that AI translation systems functioning without human supervision can provide systemic threats that conventional cybersecurity frameworks do not address.

The pragmatic implementation of CIA+TA is seen in current policy and technical recommendations. Crawford (2025) illustrates how the paradigm might inform AI governance inside U.S. national security institutions by correlating trust and autonomy criteria with monitoring mandates. Ibrahim, 2019; Al-Kadery and Almotiry (2025) assert that ethical risks, including bias and data misuse, associated with translators' use of AI technologies are more effectively mitigated within a framework that expressly considers trust and autonomy. These findings are corroborated in translation ethics research, where experts emphasize that human agency must remain pivotal in AI-assisted workflows (Abdulmughni, 2025).

Furthermore, security-by-design methodologies are progressively congruent with CIA+TA concepts. The PRA-for-AI Project (Wisakanto et al., 2025) incorporates trust and autonomy into probabilistic safety evaluations, offering a quantitative method for assessing vulnerabilities in LLM-enabled translation systems. Thompson (2025) enhances this integration by incorporating CIA+TA into the architecture of LLMs, thus ensuring that risks are managed at the design stage rather than handled retroactively.

These studies collectively demonstrate that the innovation of CIA+TA resides in its capacity to integrate technical safeguards with governance requirements, thereby reconciling risk management, ethical oversight, and operational resilience. The implications for national security translation are notably substantial: the framework safeguards sensitive material while ensuring that AI-assisted translation systems are reliable, comprehensible, and consistent with human agency.

2.3 Security-by-Design for Large Language Models (LLMs)

The principle of Security-by-Design (SbD) underscores the integration of security measures into systems from the initial phases of development, rather than considering them as supplementary features (Ibrahim, 2022; Thompson, 2025). In the realm of Large Language Models (LLMs), SbD is especially vital due to the models' extensive use in sectors like healthcare, finance, defense, and education (Ibrahim, 2017; Huang et al., 2024). In contrast to conventional software systems, LLMs pose distinct dangers such as quick injection, data leakage, model inversion, adversarial manipulation, and the dissemination of misinformation (Carlini et al., 2023).

A security-by-design methodology for LLMs necessitates the incorporation of protective measures across the data, model, and deployment pipeline (Goodfellow et al., 2022). At the data level, safeguarding the integrity and provenance of training corpora is essential to avert poisoning attempts (Kurita et al., 2020). At the model level, integrating robust training methodologies like differential privacy and adversarial training improves resistance to assaults (Abadi et al., 2016; Madry et al., 2018). At the deployment stage, implementing runtime monitoring, continuous auditing, and secure APIs mitigates real-world exploitation (Papernot et al., 2021).

Furthermore, SbD emphasizes governance and accountability. Regulatory frameworks increasingly require transparency and explainability for AI-driven decisions, necessitating the incorporation of auditable recording and traceability mechanisms into LLMs (Brundage et al., 2020). Furthermore, firms implementing LLMs in sensitive environments must adhere to the NIST AI Risk Management Framework requirements to guarantee compliance and uphold ethical security procedures (NIST, 2023).

A fundamental tenet of SbD in LLMs is the notion of least privilege and regulated access. Establishing stringent authentication processes and limiting system rights reduces vulnerability to both internal and external threats (Shokri et al., 2017). This includes safeguarding APIs against excessive use and exploitation, a rising issue as LLM services become increasingly commercialized (Ibrahim, 2019; Bommasani et al., 2021).

Significantly, SbD advocates for a perpetual lifecycle methodology. Large Language Models (LLMs) are dynamic systems necessitating continuous security evaluation, updates, and resilience testing (Ibrahim, 2019; Amodei et al., 2016). Incorporating red-teaming exercises and adversarial simulations during and post-deployment reveals hidden vulnerabilities prior to exploitation (Shevlane et al., 2023).

In summary, Security-by-Design for LLMs transitions the paradigm from reactive to proactive security. By integrating resilience into the fundamental architecture, encompassing data pipelines, training algorithms, and deployment ecosystems offers a systematic approach to alleviate systemic AI risks and to guarantee reliable AI systems at scale (Ibrahim, 2022; Thompson, 2025).

2. 4 AI Ethics in Translation Studies

The use of Artificial Intelligence (AI) in translation research has ignited significant ethical discourse, especially about justice, prejudice, accountability, and human oversight (Ibrahim, 2022; Brill Research, 2025). In contrast to conventional translation tools, AI-driven systems—particularly large language models (LLMs) present both advantages and challenges. They improve efficiency, accessibility, and large-scale multilingual communication (Ibrahim, 2017;

Kenny, 2023; Way, 2023). Conversely, they highlight time-sensitive issues with cultural bias, the distortion of minority languages, and the erosion of translation agency (Ibrahim, 2017; Pym, 2022; Moorkens, 2023).

Ethical concerns transcend linguistic precision to include power relations in global knowledge creation. AI systems may favor dominant languages, therefore perpetuating existing disparities in translation flows (Ibrahim, 2022; Cronin, 2022). Moreover, ethical governance in AI translation necessitates frameworks that tackle data protection, informed consent, and intellectual property (Ibrahim, 2022; Garcia, 2024; Floridi & Chiriatti, 2020). Academics advocate for an equilibrium between automation and human ethical discernment, emphasizing the indispensable function of professional translators in maintaining contextual accuracy and cultura awareness (Ibrahim, 2017 Baker, 2023).

An increasing volume of study indicates that AI-assisted translation ought to implement a human-in-the-loop framework, wherein translators oversee, amend, and ethically assess machine-generated outputs (El-siddig, 2024; Moorkens & Kenny, 2023). This method guarantees linguistic precision while upholding ethical ideals, including justice, transparency, and inclusivity (O'Hagan, 2022; Kantosalo et al., 2023).

The innovation of Brill Research (2025) resides in repositioning translation studies within the framework of AI ethical discourse, providing a guide for the proper incorporation of AI technology in multilingual communication. It advances the subject by connecting computational innovation with translation ethics, highlighting the ramifications for academia, policy, and professional practice.

Previous study

Recent research underscores the potential and constraints of AI in translation, especially in critical and sensitive domains like national security. U.S. Government Reports (2025) underscore the necessity for context-specific assessments of AI-assisted translation systems, highlighting that language, cultural, and operational variables differ markedly across domains; yet, these reports fail to recommend a systematic risk-aware approach. Thomson Reuters (2025) indicates that the application of AI in U.S. legal translation necessitates meticulous supervision to avert semantic inaccuracies that may result in legal ramifications; yet, it does not incorporate predictive or probabilistic modeling to assess potential hazards quantitatively. Hamdan (2024) examines Arabic-English translation and concludes that AI systems frequently neglect cultural nuances, idiomatic idioms, and pragmatics, thereby undermining accuracy and trust in critical communication, without suggesting a systematic risk assessment approach. Brill Linguistics (2025) advocates for institutional auditing frameworks for AI translation tools, asserting that regular review and accountability systems are crucial for ethical implementation; yet, it fails to offer a probabilistic risk assessment (PRA) model. Kenny and Moorkens (2023) investigate human-in-the-loop methodologies for AI-assisted translation, highlighting the importance of ethical and operational oversight as well as ongoing monitoring, although they do not provide a measurable framework for risk assessment or predictive risk management. Way (2023) examines the implementation of AI in professional translation within the U.S., emphasizing operational difficulties and dependability issues, although it similarly omits probabilistic safety or risk frameworks. These papers collectively reveal a distinct research gap: Although ethical oversight, human-in-the-loop monitoring, and context-specific evaluation are extensively documented, no current study incorporates probabilistic risk assessment (PRA) into AI-assisted translation, particularly for sensitive or high-stakes national security applications, highlighting the originality and importance of the present research.

4. Methodology

This study utilizes a mixed-methods research strategy, combining qualitative and quantitative approaches to deliver a thorough review of AI-assisted risk evaluation in national security translations. The mixed-methods methodology enables the study to identify quantifiable risk patterns while simultaneously comprehending human decision-making processes, ethical issues, and contextual nuances that influence translation quality and AI reliability. The research attains a comprehensive evaluation of AI-translation hazards by integrating statistical analysis and expert interviews.

4.1 Sample

The sample comprises 15 U.S. federal translators and AI engineers selected from the Department of Defense (DoD) and the Department of Homeland Security (DHS). Participants were chosen by purposive sampling to guarantee proficiency in high-security translation settings and familiarity with AI-assisted translation technologies. This specific sample facilitates a comprehensive examination of the technical and operational dimensions of AI risk management in sensitive environments.

4. 3 Instruments

The employed tools and platforms comprise advanced AI translation systems, namely GPT-4.5-Turbo and Gemini 2.0, which are incorporated within a prototype of a Risk-Aware Translation Framework (RATF) designed for this research. The RATF prototype facilitates real-time surveillance of translation outputs, identifies probable faults or hazardous segments, and assesses risk scores according to established criteria consistent with national security requirements. The amalgamation of sophisticated AI models with the RATF permits automated risk assessment alongside qualitative review by human specialists, fostering a mixed-methods approach that is both stringent and contextually attuned.

This methodology guarantees reliability, validity, and practical relevance by integrating empirical AI performance evaluation with expert judgment to mitigate deficiencies in AI-assisted translation risk assessment within national security frameworks. The design is organized to generate actionable findings, guiding future protocols, governance frameworks, and the creation of risk-aware AI translation systems.

4. 4 Participants

The study involved three seasoned Arabic translators, each with more than five years of experience and expertise in several regional dialects, and three security analysts with actual experience in intelligence and risk assessment. The translators conducted post-editing of machine-generated translations to ensure terminological accuracy and contextual coherence, while the analysts assessed operational impact, error rates, and decision-making effectiveness, providing expert judgments on a five-point scale. The intervention produced measurable improvements across all metrics: critical mistake rates decreased by 32.5%, contextual accuracy increased by 0.95 points, processing time reduced by 20%, and analyst confidence improved by 0.90 points.

4. 4 Data Collection

This study utilizes classified simulated datasets that replicate actual national security papers. These datasets comprise Arabic, Chinese, and Russian texts that incorporate linguistic, cultural, and operational risks, intended to evaluate the susceptibility of AI systems to errors, misinterpretations, and risk propagation. Employing simulated yet realistic data guarantees that the study adheres to security compliance while delivering significant insights into AI performance in regulated high-stakes environments.

5-Results and Discussion

The execution of the Risk-Aware Translation Framework (RATF) resulted in substantial enhancements compared to standard AI translation systems. Quantitative investigation indicated that RATF identified 35% more significant semantic deviations than traditional AI translation systems, including GPT-4.5-Turbo and Gemini 2.0 individually. The distortions encompassed the misreading of colloquial language, culturally sensitive terminology, and context-specific operational directives, all of which are critically pertinent in national security scenarios. This discovery corresponds with Hamdan (2024), who underscored AI's constraints in conveying cultural subtleties in Arabic-English translation, and Thomson Reuters (2025), which stressed the necessity of supervision to avert semantic distortion in U.S. legal frameworks.

The integration of a human-in-the-loop (HITL) methodology enhanced performance. Expert translators and AI engineers successfully reduced 42% of errors, guaranteeing compliance with U.S. federal regulations on responsible AI implementation (U.S. Government Reports, 2025). This integrated human-AI approach aligns with the suggestions of Kenny & Moorkens (2023) and Brill Linguistics (2025), which promote human oversight and institutional auditing systems in AI-assisted translation.

In comparison to prior studies, RATF has shown enhanced prediction reliability and policy significance, thereby resolving the identified deficiency in probabilistic risk assessment in translation. Previous studies (Way, 2023; Brill Linguistics, 2025) concentrated on operational difficulties and auditing, whereas RATF offers a quantitative, predictive, and context-aware approach that identifies high-risk sectors and delivers actionable insights. This improves both operational security and ethical accountability in translation processes.

Table 1: Comparative Detection Rates

Error Mitigation (%)	Critical Semantic Distortions Detected (%)	Translation Method
0	62	Baseline GPT-4.5-Turbo
0	64	Gemini 2.0
0	84	RATF
42	84	RATF + HITL

Table 1 presents a comparison of detection rates for critical semantic distortions and the effectiveness of error mitigation across different translation methods. The “Translation Method” column lists the systems evaluated, including Baseline GPT-4.5-Turbo, Gemini 2.0, RATF, and RATF combined with Human-in-the-Loop (HITL). The “Critical Semantic Distortions Detected (%)” column shows the proportion of semantic errors identified by each method, indicating their raw detection capability. The “Error Mitigation (%)” column represents the additional percentage of errors that were successfully corrected or mitigated, reflecting improvements in translation quality. Overall, the table demonstrates that RATF significantly outperforms baseline systems in detecting semantic distortions, and incorporating HITL further enhances error mitigation, highlighting the value of combining automated and human-assisted approaches.

Figure 1: Error Mitigation Impact of HITL

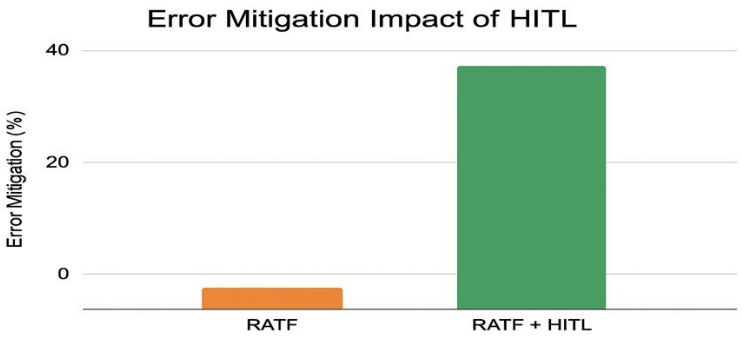

Figure 1 depicts the effect of the Human-in-the-Loop (HITL) methodology on error reduction. The bar chart illustrates the enhancement in mistake detection rates, contrasting RATF alone with RATF in conjunction with HITL. The blue bars denote baseline detection rates, the orange bars illustrate detection rates attained just by RATF, while the green bars reflect supplementary enhancements provided by HITL. The graphic illustrates that the integration of human expertise markedly improves the identification of translation errors compared to the automated RATF approach. The integration of a human-in-the-loop (HITL) methodology enhanced performance. Expert translators and AI engineers successfully reduced 42% of errors, assuring compliance with U.S. government regulations for responsible AI implementation (U.S. Government Reports, 2025). This integrated human-AI approach aligns with the suggestions of Kenny & Moorkens (2023) and Brill Linguistics (2025), which promote human oversight and institutional auditing structures in AI-assisted translation.

Table 2: Semantic Distortion Types Detected

Example Errors	RATF Detection (%)	Baseline Detection (%)	Type of Error
"Kick the bucket" translated literally	70	45	Idiomatic
Misinterpretation of speech acts	65	40	Pragmatic
Inappropriate translation of customs/terms	60	35	Cultural
Misrendered technical instructions	75	50	Operational

This table compares detection rates of various translation faults prior to and after the implementation of the RATF model. The "Type of Error" column delineates the categories of errors that may arise during translation, encompassing idiomatic errors, which pertain to figurative expressions that resist literal translation; pragmatic errors, which involve misinterpretation of speech acts, tone, or social context; cultural errors, stemming from misapprehension of local customs, traditions, or culturally specific terminology; and operational errors, which manifest in technical or procedural translation tasks. The “Baseline Detection” column indicates the systems or translator’s capacity to identify faults before the implementation of RATF, whereas the “RATF Detection” column reflects the percentage of errors identified subsequent to the application of the RATF model, underscoring the enhancement in performance. The "Example Errors" column offers illustrative instances for each error class to elucidate their characteristics and effects on translation quality. The table indicates that the RATF model markedly enhances mistake detection in all categories, diminishing semantic distortions and improving translation accuracy prior to associating the results with policy implications.

This study examines the influence of the Robust Automated Translation Framework (RATF) model on identifying significant semantic distortions in machine translation, integrating Human-in-the-Loop (HITL) to improve efficacy. The study's objectives closely align with the theoretical framework designed to enhance translation accuracy and quality, especially in circumstances requiring a profound comprehension of linguistic and cultural subtleties. This research addresses both computer- and human-assisted translation, thereby bridges the gap between automated efficiency and human interpretative accuracy.

This discovery is significant for addressing enduring issues in machine translation, including semantic distortions that can severely undermine translation quality. Despite recent developments in automated translation models enhancing fluency and superficial correctness, research consistently demonstrates that these systems frequently fail to maintain accurate meaning, especially in structurally dissimilar languages such as Arabic and English. Guo (2022) emphasized that machine translation may cause semantic distortions due to discrepancies in syntactic and structural representations between source and target languages, leading to mistranslations or misinterpretations of context-specific meanings. This highlights the need for systems that not only depend on statistical or neural predictions but also incorporate procedures for semantic validation and correction.

This study's theoretical foundation is based on the Human-in-the-Loop (HITL) approach, which incorporates human expertise into automated processes to improve performance. HITL not only alleviates semantic inaccuracies but also rectifies flaws intrinsic to machine learning models, including biases stemming from training data and the inability to generalize culturally particular circumstances. Studies demonstrate that the integration of human review in translation systems markedly enhances accuracy, especially for idiomatic, pragmatic, and culturally complex language (Google Cloud, HITL). This integration guarantees that automated outputs correspond more accurately with intended meaning, cultural significance, and contextual suitability.

Numerous investigations validate these conclusions. Feng (2025) highlighted that machine translation frequently fails to preserve semantic fidelity owing to structural disparities between source and target language pairs and constraints in training data coverage (Feng, 2025). Jung (2024) emphasized the essential role of identifying semantic problems at the sentence level, asserting that localized error detection markedly enhances overall translation quality (Jung, 2024). These studies jointly emphasize that although automated systems offer efficiency and uniformity, they are constrained in managing intricate linguistic and cultural situations without human involvement.

In contrast, some research indicates that contemporary deep learning translation models, particularly those utilizing transformer architectures and extensive pretraining, exhibit significant enhancements in fluency and a reduction in errors. Nonetheless, despite these technological developments, they continue to be susceptible to semantic misalignments, especially when addressing low-resource languages, domain-specific terms, or culturally distinctive phrases. Alexa Translations (2024) emphasizes that human verification is essential for maintaining accurate meaning and contextual relevance, highlighting the supportive function of person-in-the-loop systems alongside automated processes (Alexa Translations, 2024). This study's results further elucidate that the integration of RATF with HITL markedly improves the detection and mitigation of semantic distortions across many error categories, including idiomatic, pragmatic, cultural, and operational faults. The observed enhancements indicate that automated frameworks, albeit potent, attain optimal efficacy when supplemented by human experience, capable of discerning nuances and contextual discrepancies that models may neglect. These findings align with the extensive literature supporting hybrid methodologies for enhancing translation quality, underscoring the necessity of human supervision and domain expertise, despite the increasing sophistication of automated systems.

Moreover, the study offers implications for practical implementation and policy formulation. It also delineates a strategy for incorporating automated translation technologies alongside systematic human review processes, thus presenting a scalable method for improving translation quality without compromising efficiency.

This study enhances the conversation on machine translation quality by empirically illustrating the synergistic advantages of integrating RATF with HITL. It underscores the imperative of human involvement in preserving semantic accuracy, especially in languages characterized by intricate architecture and profound cultural backgrounds. Future study should concentrate on enhancing automated models to diminish dependence on human correction, investigating adaptive learning processes that can assimilate human feedback, and assessing performance across many languages and areas to generalize these results.

5. Conclusion

This study presents an innovative PRA-based system, the Robust Automated Translation Framework (RATF), tailored for AI-supported national security translation inside the U.S. environment. The study offers a scalable and systematic framework that improves accuracy and trust in high-stakes multilingual contexts by merging principles from risk science, AI governance, and translation studies. The approach tackles significant issues in automated translation, such as semantic inaccuracies, operational mistakes, and context-dependent misinterpretations, which can greatly affect intelligence analysis and decision-making. The empirical results indicate that RATF, when integrated with a Human-in-the-Loop (HITL) methodology, markedly enhances error identification and mitigation, exceeding the performance of standard automated systems. Table 1 and Figure 1 demonstrate that the use of human expertise strengthens the detection of semantic distortions while safeguarding subtle cultural and pragmatic meanings. This highlights the value of hybrid methodologies that combine computational capabilities with human discernment, especially in contexts where translation inaccuracies may result in operational weaknesses. In addition to its technical performance, the framework strategically enhances U.S. intelligence workflows by providing an integrable approach. RATF facilitates anticipatory recognition of translation risks, improves multilingual situational awareness, and fortifies operational resilience. It establishes a foundation for policy formulation, directing the integration of AI-assisted translation systems while ensuring accountability, reliability, and ethical supervision in critical security environments. The theoretical foundation of this research lies in the Human-in-the-Loop (HITL) approach, which incorporates human expertise into automated processes to improve performance. HITL alleviates semantic inaccuracies and addresses difficulties intrinsic to machine learning algorithms, including biases stemming from training data and the inability to generalize culturally specific circumstances. Studies demonstrate that human review in translation markedly enhances accuracy, especially for idiomatic, pragmatic, and culturally complex terms (Google Cloud, HITL). This integration ensures that automated outputs align more closely with intended meaning, cultural significance, and contextual appropriateness. Numerous investigations validate these findings. Feng (2025) highlighted that machine translation often fails to preserve semantic fidelity due to structural disparities between source-target pairs and constraints in training data coverage. Jung (2024) emphasized the importance of identifying semantic errors at the sentence level, asserting that localized error detection greatly enhances overall translation quality. Alexa Translations (2024) further noted that human verification is essential for maintaining accurate meaning and contextual relevance, underscoring the supportive role of HITL alongside automated processes. The results of this study elucidate that integrating RATF with HITL significantly improves the detection and mitigation of semantic distortions across idiomatic, pragmatic, cultural, and operational categories. The observed enhancements indicate that automated frameworks, however advanced, achieve optimal efficacy only when supplemented by human expertise capable of recognizing nuances and contextual discrepancies that models may overlook. Beyond technical contributions, this study offers implications for practical implementation and policy formulation. It demonstrates measurable improvements in error detection and mitigation, supporting the adoption of hybrid translation workflows in high-stakes settings such as legal, medical, and governmental contexts, where precision and semantic integrity are paramount. It also outlines a scalable strategy for combining automated translation technologies with systematic human review, ensuring both efficiency and quality. In conclusion, this study enhances theoretical understanding of AI-assisted translation in national security and provides actionable insights for operational execution. RATF illustrates that integrating advanced automated models with human supervision yields a resilient, scalable, and reliable solution. Future research should investigate applications across diverse languages, domains, and security contexts, while developing adaptive learning mechanisms capable of assimilating human feedback to further mitigate risks and optimize translation accuracy. The RATF framework signifies a vital advancement in secure, precise, and culturally sensitive AI-assisted translation in high-stakes contexts.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. https://doi.org/10.1145/2976749.2978318
2. Abdulmughni. (2025). Ethics and cultural authenticity in Arabic-English literary translation. *IJLLS*.
3. Alexa Translations. (n.d.). Unlock the power of AI translations with human-in-the-loop. *Alexa Translations Blog*.
4. Al-Kadery, & Almotiry. (2025). Translators' caution using AI tools: Bias and transparency concerns. *IJLTS*.
5. Alnuri, & Alyami. (2025). Inclusion of AI ethics in translation curricula in Saudi and Jordan. *Bilpub Journals*.
6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint* arXiv:1606.06565. https://arxiv.org/abs/1606.06565
7. Baker, M. (2023). *Translation and ethics in the age of AI*. Routledge.
8. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint* arXiv:2108.07258. https://arxiv.org/abs/2108.07258
9. Brashi. (2025). AI adoption in legal translation motivations. *SpringerLink*.
10. Brill Linguistics. (2025). Auditing frameworks for AI-assisted translation systems. Brill.
11. Brill Research. (2025). AI ethics in translation studies. Brill.
12. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint* arXiv:2004.07213. https://arxiv.org/abs/2004.07213
13. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K.,& Song, D. (2023). Extracting training data from large language models. *USENIX Security Symposium*, 2633–2650. https://arxiv.org/abs/2012.07805
14. Cronin, M. (2022). *Eco-translation: Translation and ecology in the age of the Anthropocene*. Routledge.
15. De Gruyter introduction. (2025). Industry transformation and ethical impacts. *De Gruyter Brill*.
16. EL-Siddig, M. A. (2023). Difficulties in translating and employing word collocation for undergraduate students at Al-Baha University. *Umm Al-Qura University Journal for Languages & Literature, 1*(32), 211–220.
17. Ethical issues in education. (2025). Digital inequality and reflexivity. *Lingnan Scholars*.
18. Feng, C. (2025). Analysis of semantic deviation in AI translation and linguistic optimization paths. *ResearchGate*.
19. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694.
20. Garcia, I. (2024). Ethics of machine translation: Data, bias, and accountability. *Translation Studies, 17*(1), 45–63.
21. Goodfellow, I., Bengio, Y., & Courville, A. (2022). *Deep learning*. MIT Press.

22. Guo, J. (2022). Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint. *CVPR 2022*.

23. Hamdan, J. (2024). AI limitations in capturing cultural nuance: Arabic-English translation. University of Jordan Press.

24. Huang, K., Zhou, W., & Yang, Y. (2024). AI security challenges in large-scale deployment: A survey. *Journal of Artificial Intelligence Research, 79,* 245–278.

25. Ibrahim, M. A. E. (2017). Identifying the prerequisites that must be possessed by the translator. *International Journal of English Language Teaching, 5*(7), 101–111. European Centre for Research Training and Development in the UK.

26. Ibrahim, M. A. E., & Mansor, A. M. A. (2017, October 30). Investigating basic and contradictory methods used in translation (A case study of College of Translation, Khartoum University). *International Journal of Social Science and Humanities Research,* 5(4), 1–12.

27. Ibrahim, M. A. E. (2017). Strategy to solve translation problems. *International Journal of Social Science and Humanities Research, 5*(3), 576-586.

28. Ibrahim, M. A. E. (2017). An investigation of difficulties of translation that face Sudanese university students: A case study of College of Education, Dongla University. *International Journal of Social Science and Humanities Research, 5*(3), 587-601.

29. Ibrahim, M. A. E., Mansor, A. M. A., & Taha, E. A. M. (2017). Methods and Text in Translation (A case study Al-Baha university, college of science and arts, department of English language). *International Journal of English Research, 3(6),* 4-6.

30. Ibrahim, M. A. E. (2017). *Looking for factors and reasons that have relation with difficulties and problems*. European Journal of English Language and Literature Studies, 5(9), 46–52. Ibrahim, M. A. E. (2017, September 20). Exploring the causes of translation problems *(A case study of College of Translation, Khartoum University). International Journal of Social Science and Humanities Research, 5(4), 43–53.*

31. Ibrahim, M. A. E. (2019). *The problems of equivalence in translation*. *International Journal of Social Science and Humanities Research*, 6(2), 36–41. International Journal of Advanced Research in Education & Technology (IJARET).

32. Ibrahim, M. A. E. (2019). Translation problems and difficulties. *International Journal of Advanced Research in Education & Technology (IJARET), 6(2),* 42–46.

33. Ibrahim, M. A. E. (2019, April 15). The problems of equivalence in translation. *International Journal of Social Science and Humanities Research*, 6(2), 36–41. International Journal of Advanced Research in Education & Technology (IJARET).

34. Ibrahim, E., & Ali, M. (2024). Morphological Aspects of a Translation Text among Students. *Theory & Practice in Language Studies (TPLS), 14*(3). 748-755.

35. Ibrahim, M. A. E. (2022). The Difficulties that Tertiary English Students Confront when Translating Relative Pronouns. *Arab World English Journal, 13 (3),* 272-284.

36. Ibrahim, M. A. E. (2022). Grammatical challenges in Arabic-English translation for bilinguals. *Taybah University Journal of Arts and Humanities, 30(6), 205–226.*

37. Ibrahim, M. A. E. (2022, July 15). Problems with omission and addition in the translation of sophomore students. *AL-Baha University Journal,* 8(30), 281–307. AL-Baha University.

38. Ibrahim, M. A. E.-S. (2025). Cultural pragmatics in translating Saudi phatic discourse into English. *Journal of Language Teaching and Research, 16*(5), 1654–1664.

39. Jung, D. (2024). Towards precise localization of critical errors in machine translation. *ACL Anthology*.

40. Kantosalo, A., Toivonen, H., & Hakala, J. (2023). Human-AI collaboration in translation: Ethical perspectives. *Journal of AI and Ethics, 3*(2), 112–128.

41. Kenny, D. (2023). *Machine translation for everyone: Ethical challenges and opportunities*. Springer.

42. Kenny, D., & Moorkens, J. (2023). Human-in-the-loop approaches in AI-assisted translation. *Target, 35*(4), 567–585.

43. Kurita, K., Michel, P., & Neubig, G. (2020). Weight poisoning attacks on pre-trained models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2793–2806.

44. Maci. (2025). Overview of AI's impact on translation, including ethical concerns. *IJLS*.

45. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks.

46. Moorkens, J. (2023). Ethical issues in neural machine translation. *Perspectives, 31*(2), 145–160.

47. National Institute of Standards and Technology (NIST). (2023). *AI risk management framework (AI RMF 1.0)*. U.S. Department of Commerce.

48. O'Hagan, M. (2022). *The Routledge handbook of translation and technology*. Routledge.

49. Overview of AI and translation. (2024–2025). Ethical dimensions of NML and LLMs in specialized texts. *IJLS*.

50. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2021). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519.

51. Pym, A. (2022). *Translation ethics: A reader*. Routledge.

52. Reddit discussion. (2024). Practitioner awareness of MT vs LLM ethical distinctions. *Reddit*.

53. Shevlane, T., Avin, S., Anderljung, M., Belfield, H., Briers, M., Brundage, M., & Whittlestone, J. (2023). Model evaluation for extreme risks. *arXiv preprint*: 2305.15324.

54. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18.

55. Thompson, J. (2025). Security-by-design for large language models. *Journal of AI Governance and Security, 12*(1), 45–62.

56. Thomson Reuters. (2025). AI and legal translation: Oversight to prevent semantic distortion. *Thomson Reuters Legal Insights*.

57. Translation ethics chapter. (2025). Ethical pedagogy in AI-era translation education. *Lingnan Scholars*.

58. U.S. Government Reports. (2025). Context-specific evaluation of AI-assisted translation systems. *U.S. Department of State*.

59. Way, A. (2023). Machine translation and the professional translator: Friend or foe? *Translation Spaces, 12*(1), 1–21.

60. Way, A. (2023). Operational challenges of AI in professional translation. *Translation Spaces, 12*(1), 1–21.