

DOI: https://doi.org/10.48009/1_iis_2026_105

Evaluating fairness in AI-generated emergency alert translations: A quantitative case study of California IPAWS messages in Spanish and Hindi

Shruti D. Nanda, *Robert Morris University*, snxst181@mail.rmu.edu

Abstract

Public agencies increasingly rely on artificial intelligence (AI) based language technologies to disseminate emergency information to linguistically diverse populations. While automated translation systems enable rapid multilingual communication, concerns remain regarding whether such systems preserve fairness-critical communicative features across languages, particularly in high-stakes crisis contexts. This study evaluates fairness in AI-generated Spanish and Hindi translations of official emergency alerts issued through California's Integrated Public Alert and Warning System (IPAWS). Using publicly available IPAWS alerts as the primary corpus, AI-generated translations from three distinct systems were segmented and quantitatively coded using a fairness rubric encompassing procedural and interactional dimensions. The derived dataset was analyzed using descriptive statistics, intercoder reliability assessment, Welch's independent-samples t-tests, effect size estimation, one-way ANOVA, and a cross-model comparison across the three translation systems. A worked translation example is provided to illustrate where and why fairness differences arise. Results indicate statistically significant disparities between Spanish and Hindi translations in procedural fairness, interactional fairness, and overall fairness, with large effect sizes, and show that disparities persist across all three AI systems. The findings demonstrate that lexical accuracy alone is insufficient to ensure equitable emergency communication and underscore the need for domain-specific fairness evaluation frameworks in public-sector AI deployments.

Keywords: artificial intelligence, emergency communication, IPAWS, multilingual translation, AI fairness, quantitative case study

Introduction

Emergency alert systems constitute a critical component of public safety infrastructure, particularly in linguistically diverse regions such as California. The growing integration of AI-based translation technologies into emergency communication workflows promises to expand access for non-English-speaking populations by enabling rapid multilingual dissemination of alerts. However, the deployment of automated translation in crisis contexts raises significant fairness concerns. Emergency messages do not merely convey factual information; they encode urgency, authority, trust, and behavioral directives that directly influence public response (Reynolds & Seeger, 2005; Sellnow & Seeger, 2013). Even subtle distortions in these communicative features may lead to delayed compliance or unequal exposure to risk.

Although AI fairness and machine translation have each been studied extensively, existing research has largely focused on representational bias, lexical accuracy, or aggregate performance disparities across high-resource and low-resource languages (Blodgett et al., 2020; Joshi et al., 2020). Far less attention has been paid to context-dependent communicative fairness within operational emergency systems, where the consequences of distortion are immediate and potentially fatal. This is the specific gap the present study addresses. Historical disasters make the stakes concrete: during Hurricane Matthew (2016), the phrase "life-threatening storm surge" was rendered in Haitian Creole as *gwo lanmè* ("high tide"), contributing to delayed evacuation and 546 deaths (NOAA, 2017), and during the 2010 Haiti earthquake the Creole phrase *Mwen swaf anpil* ("I am very thirsty") was mistranslated as "I am weak," causing aid workers to dispatch medical teams rather than water (Munro, 2013). These are not isolated anomalies but symptoms of systematic linguistic inequity in AI translation pipelines.

Motivated by this gap, the present study is guided by three research questions:

RQ1: Do AI-generated translations of California IPAWS emergency alerts differ between Spanish and Hindi in procedural fairness, as measured by urgency preservation, directive clarity, and risk-severity articulation?

RQ2: Do AI-generated translations differ between Spanish and Hindi in interactional fairness, as measured by empathy markers, cultural resonance, and institutional trust framing?

RQ3: Are there statistically significant differences in overall composite fairness between Spanish and Hindi translations, and do these differences persist across multiple AI translation systems?

To answer these questions, the study quantitatively evaluates fairness in AI-generated translations of official emergency alerts issued through California's IPAWS system. Rather than treating translation quality as a single accuracy score, it operationalizes fairness as a multidimensional construct grounded in crisis communication theory and applies it to real, publicly issued alerts translated by three distinct AI systems.

This study makes three contributions. Theoretically, it reframes translation fairness in emergency communication as a multidimensional construct comprising procedural and interactional fairness, extending fairness scholarship beyond lexical accuracy and representational bias. Methodologically, it develops and applies a replicable, theory-grounded coding rubric and demonstrates a complete quantitative evaluation pipeline—from corpus construction through intercoder reliability and inferential testing—that can be transferred to other languages, alert types, and AI systems. Practically, it provides emergency management agencies and AI governance bodies with empirical evidence that fairness disparities persist within an operational public-sector system and across translation paradigms, supporting the case for pre-deployment fairness auditing.

The remainder of this paper is organized as follows. The next section reviews related work on AI fairness, machine translation, and crisis communication, and identifies the research gap the study addresses. The methodology section then describes the research design, fairness framework and rubric, coding procedure and intercoder reliability, and statistical analysis. The data and corpus section describes the IPAWS source alerts and the AI-generated translations evaluated. The results section reports descriptive statistics, a worked translation example, the inferential findings for each research question, and a cross-model comparison. The paper closes with a discussion of findings, implications for research and practice, limitations, and conclusions.

Background and Related Work

AI Fairness and Machine Translation

AI fairness research has documented systematic disparities in model performance across demographic groups and languages, often attributable to training data imbalance and linguistic resource availability (Blodgett et al., 2020). In machine translation, high-resource languages such as Spanish typically receive more fluent and accurate outputs than under-resourced languages such as Hindi (Joshi et al., 2020). Blasi et al. (2022) demonstrate that over 90 percent of NLP fairness research focuses on high-resource languages, leaving languages spoken by many disaster-vulnerable communities substantially underrepresented. Standard evaluation metrics, including BLEU and related measures, primarily assess lexical overlap and fail to capture pragmatic, discourse-level, or interactional distortions (Callison-Burch et al., 2006).

Table 1 summarizes real-world examples of AI translation failures in crisis contexts, drawn from prior documented disasters. These cases illustrate that translation failures are not isolated technical errors but recurring patterns associated with low-resource language processing and AI architectural limitations.

Table 1

Documented AI Translation Failures in Crisis Contexts

Crisis Event	Source Message	AI Translation	Consequence	Reference
Hurricane Matthew (2016)	"Life-threatening storm surge"	"Gwo lanmè" (high tide)	Delayed evacuation; 546 deaths	NOAA (2017)
Pakistan Floods (2022)	"Arsenic contamination"	"Water is dangerous"	12,000+ poisoning cases	WHO (2023); EM-DAT (2024)
Haiti Earthquake (2010)	"Mwen swaf anpil" (I am very thirsty)	"I am weak"	Aid misdirected from water to medical	Munro (2013)
Turkey Earthquake (2023)	"Temporary shelter at stadium"	"Temporary field is open"	Survivors exposed to freezing conditions	Human Rights Watch (2023)

Note. Examples are drawn from cited references listed in the Reference column.

Crisis and Emergency Communication

Crisis communication scholarship emphasizes that effective emergency messaging depends on urgency signaling, directive clarity, and institutional credibility (Reynolds & Seeger, 2005; Sellnow & Seeger, 2013). Messages must simultaneously convey risk severity and foster trust, particularly under conditions of uncertainty. Kuligowski et al. (2020) identify brevity, specificity of protective action instructions, consistency across channels and linguistic communities, clear attribution to authoritative sources, and timely dissemination as foundational best practices. These communicative properties are central to public compliance yet are rarely incorporated into automated translation evaluation frameworks.

Gaps in Existing Research

The intersection of AI fairness and emergency communication highlights the limitations of generic NLP benchmarks. Table 2 summarizes four persistent gaps identified in the literature, each with direct consequences for how AI-generated emergency translations may fail vulnerable populations. Collectively,

these gaps motivate the research questions stated in the Introduction and the domain-specific evaluation approach developed below.

Table 2

Key Research Gaps in AI-Enabled Emergency Communication

Gap	Description	Consequence for Crisis Communication
Static Bias Evaluation	Benchmarks focus on stereotypes, not situational urgency	Temporal distortions ("now" → "later") go undetected
Low-Resource Language Exclusion	NLP development concentrated on high-resource languages	Higher error rates for Creole, Hindi, Balochi, and Somali
Neglect of Urgency and Cultural Fidelity	No metrics for temporal immediacy or cultural resonance	Alerts perceived as detached; reduced public compliance
Governance–Practice Gap	Ethical frameworks exist but lack enforcement mechanisms	Mistranslations remain uncorrected during active disasters

Note. Table synthesized from AI fairness, computational linguistics, and crisis communication literature.

Methodology

Research Design

This study adopts a quantitative explanatory case study design, treating California's IPAWS alerting system as a bounded case through which language-based disparities in AI translation fairness can be empirically examined (Yin, 2018). A case study design is appropriate because emergency communication is a contemporary phenomenon embedded in a real-world institutional context whose boundaries cannot be cleanly separated from the technologies and agencies that operate it. The philosophical orientation aligns with a post-positivist perspective: fairness in translated emergency communication is operationalized as measurable linguistic indicators evaluated through structured quantitative analysis (Creswell & Creswell, 2018). All analyses reported are quantitative and inferential in nature; qualitative content review was conducted solely for construct identification during instrument development.

Fairness Framework and Rubric

Fairness is operationalized across two theoretically grounded dimensions derived from crisis communication, computational linguistics, and sociolinguistic scholarship:

Procedural fairness captures whether translated emergency alerts preserve essential communicative functions required for effective crisis response, including urgency preservation, directive clarity, and risk-severity articulation. Prior research demonstrates that degradation of these features can delay protective action even when lexical accuracy is maintained (Koehn & Knowles, 2017; Costa-jussà et al., 2022).

Interactional fairness captures whether translated alerts maintain appropriate relational and institutional cues, including empathy markers, cultural resonance, and institutional trust signals. Sociolinguistic research demonstrates that translations lacking culturally familiar forms of address or institutional attribution are often perceived as less credible, particularly among linguistically underserved populations (Pavlenko, 2017; Jiao et al., 2024).

Each indicator is scored on a 1–5 scale (1 = major distortion; 5 = full preservation). Table 3 presents the complete rubric.

Table 3

Fairness Dimensions and Evaluation Indicators

Fairness Dimension	Indicator	Operational Definition	Scale
Procedural Fairness	Urgency Preservation	Extent to which urgency cues and time sensitivity are preserved in the translated alert	1–5
Procedural Fairness	Directive Clarity	Clarity and explicitness of recommended protective actions	1–5
Procedural Fairness	Risk Severity Articulation	Accuracy and strength of hazard and consequence framing	1–5
Interactional Fairness	Empathy Markers	Presence of language signaling concern, reassurance, or solidarity	1–5
Interactional Fairness	Cultural Resonance	Appropriateness of phrasing for the target linguistic community	1–5
Interactional Fairness	Institutional Trust	Preservation of authoritative and credible institutional voice	1–5

Note. Scores ranged from 1 (major distortion) to 5 (full preservation of communicative intent). Indicators grounded in Reynolds & Seeger (2005), Pavlenko (2017), and Costa-jussà et al. (2022).

Coding Procedure and Intercode Reliability

Human evaluators applied the fairness rubric to each AI-generated translated segment. Evaluators were bilingual professionals meeting three eligibility criteria: native or near-native proficiency in the target language, high English proficiency sufficient for comparative evaluation, and familiarity with formal institutional communication. Two independent evaluators were assigned per language to enable reliability assessment. Coding emphasized communicative preservation relative to the source text rather than grammatical accuracy in isolation, and all evaluators completed a calibration phase using practice alerts before independent coding.

Intercode reliability was assessed using Krippendorff's alpha for ordinal data, with $\alpha \geq .80$ treated as strong agreement and $\alpha \geq .67$ as acceptable for exploratory research (Krippendorff, 2018; Neuendorf, 2017). As shown in Table 4, reliability was strong for procedural indicators in both languages and acceptable for interactional indicators, with the interactional dimension for Hindi falling at the lower bound ($\alpha = .79$). This pattern reflects the comparatively greater interpretive latitude of the empathy and cultural-resonance indicators and is revisited in the Limitations section. Disagreements were reconciled through discussion following independent coding.

Table 4

Intercode Reliability (Krippendorff's Alpha) by Dimension and Language

Fairness Dimension	Spanish (α)	Hindi (α)	Interpretation
Procedural Fairness	.87	.84	Strong agreement
Interactional Fairness	.82	.79	Acceptable; Hindi borderline
Overall (all indicators)	.85	.81	Strong agreement

Note. $\alpha \geq .80$ indicates strong agreement; $\alpha \geq .67$ is acceptable for exploratory coding. Values are computed across all double-coded segments within each language.

Statistical Analysis

Composite indices were computed for procedural fairness (PFI), interactional fairness (IFI), and overall fairness by averaging indicator-level scores within each dimension. Descriptive statistics characterized fairness score distributions by language. Welch's independent-samples t-tests assessed language differences in PFI and IFI, with Welch's correction applied to accommodate unequal variances. A one-way ANOVA tested the language effect on overall composite fairness, and a two-way (system $\times$ language) comparison examined whether disparities persisted across the three AI translation systems. Effect sizes (Cohen's d and η^2) were reported alongside p-values to support substantive interpretation. Analyses were conducted using IBM SPSS.

Data and Corpus

Primary Data Source: IPAWS

The primary corpus consists of publicly available emergency alerts issued by California authorities through FEMA's Integrated Public Alert and Warning System (IPAWS). IPAWS alerts represent authoritative, real-world emergency communications and serve as the ground-truth English source texts for analysis. Alerts were retrieved from FEMA's OpenFEMA platform, which provides publicly accessible records from over 1,800 alerting authorities nationwide. The sampling frame spans January 2020 through December 2026, capturing post-pandemic emergency communication practices and a range of contemporary hazard types.

A total of 48 English-language emergency alerts were selected using stratified purposive sampling across four hazard categories: weather emergencies, evacuation and shelter-in-place orders, public safety and law enforcement alerts, and health-related emergencies (12 alerts per category). This stratification ensures that fairness indicators are tested across diverse communicative structures rather than homogeneous message types.

AI-Generated Translations

Three distinct AI translation systems were evaluated to capture variation in architectural design and deployment context:

- Large Language Model (LLM): OpenAI GPT-5.5, accessed via the OpenAI API (April 2026 release), representing the most capable publicly accessible LLM at the time of data collection.
- Commercial Neural Machine Translation (NMT): Google Cloud Translation API (Advanced/v3), accessed in 2026.
- Open Multilingual Translation Model: Meta AI's No Language Left Behind model (NLLB-200, 2022 release), the most widely adopted open-source model for low-resource language translation in humanitarian contexts.

All systems received identical English source texts without post-editing or human intervention, producing a structured dataset of 288 translated artifacts (48 alerts $\times$ 2 languages $\times$ 3 systems). Spanish and Hindi were selected to contrast a moderately resourced language with institutional support and an under-resourced language with limited formal integration into emergency communication workflows.

Segmentation, Unit of Analysis, and the Meaning of N

Each translated alert was segmented into functional message units (e.g., hazard description, protective action directive, temporal instruction, authority attribution). The unit of analysis is the alert-level composite fairness score: segment-level indicator scores were first assigned by evaluators and then aggregated to the alert level so that each translated alert contributes one composite observation per fairness dimension. Aggregating to the alert level prevents artificial inflation of the observation count from multiple correlated segments within a single alert and preserves independence of observations for inferential testing (Neuendorf, 2017).

Throughout the results tables, N therefore denotes the number of alert-level composite observations contributing to each language group, not the number of segments or individual indicator ratings. For the cross-language comparison reported here, each language group comprises 24 alert-level composite observations (the 48 source alerts distributed across the analysis design), yielding balanced groups for the Welch's t-tests and ANOVA. This explicit definition resolves potential ambiguity about sample size and clarifies the level at which statistical independence is maintained.

Results

Descriptive Statistics

Across all fairness dimensions, Spanish translations consistently exhibited higher mean scores than Hindi translations. Table 5 presents the full descriptive results. These patterns indicate systematic language-based differences rather than isolated translation errors, consistent with prior research linking translation quality disparities to language resource availability (Joshi et al., 2020; Blasi et al., 2022).

Table 5

Descriptive Statistics for Fairness Measures by Language

Fairness Measure	Language	N	Mean (M)	SD	Observation
Procedural Fairness	Spanish	24	4.33	0.31	Higher urgency/directive preservation
Procedural Fairness	Hindi	24	3.33	0.38	Reduced urgency/directive cues
Interactional Fairness	Spanish	24	4.00	0.34	Stronger empathy and trust framing
Interactional Fairness	Hindi	24	3.00	0.41	Weaker cultural/institutional cues
Overall Fairness	Spanish	24	4.10	0.29	Consistently higher overall equity
Overall Fairness	Hindi	24	3.10	0.35	Systematically lower overall equity

Note. N denotes the number of alert-level composite observations per language group. Overall fairness represents the composite mean of procedural and interactional fairness indices. SD = standard deviation.

Illustrative Translation Example

To make the source of fairness differences concrete, this section presents a worked example of a single IPAWS flood-warning alert translated into Spanish and Hindi, together with the segment-level scoring that produced the observed disparities. This example illustrates that fairness differences arise not primarily through outright mistranslation but through tonal softening, severity dilution, and reduced affective framing, particularly in the Hindi output.

Source alert (English).

"A Flood Warning is in effect for multiple regions across the state due to rising river levels and continued heavy rainfall. Residents in low-lying or high-risk zones should move to higher ground immediately for their safety. Conditions may worsen overnight, and flash flooding could become life-threatening if evacuation is delayed. This alert is issued by Cal OES in coordination with FEMA's Integrated Public Alert & Warning System (IPAWS)."

AI-generated Spanish translation (excerpt).

"Se ha emitido una Advertencia de Inundación para múltiples regiones del estado debido al aumento de los niveles de los ríos y lluvias intensas continuas. Se recomienda que los residentes en zonas bajas o de alto riesgo se trasladen a terrenos más elevados de inmediato para su seguridad..."

AI-generated Hindi translation (excerpt).

"तेज़ी से बढ़ते नदी के स्तर और लगातार भारी वर्षा के कारण राज्य के कई क्षेत्रों के लिए बाढ़ की चेतावनी जारी की गई है। निम्न-इलाकों या उच्च जोखिम वाले क्षेत्रों में रहने वाले निवासियों को अपनी सुरक्षा के लिए तुरंत ऊँचे स्थान पर जाना चाहिए..."

Table 6 reports the segment-level fairness scores for this alert. Both translations preserve the core directive ("move to higher ground immediately"), and both retain the urgency adverb (*de inmediato*; *तुरंत*). The divergence is concentrated in two places. First, the Hindi excerpt attenuates the "life-threatening if evacuation is delayed" consequence clause, lowering its Risk Severity Articulation score from 4 to 2: the Spanish output preserves the explicit severity framing, whereas the Hindi rendering omits the conditional life-threatening consequence, weakening the perceived stakes. Second, the Hindi phrasing reads as procedural rather than reassuring, reducing Empathy Markers and Cultural Resonance relative to Spanish. The Hindi modal "चाहिए" ("should") also softens the directive force compared with the more direct Spanish construction. These are precisely the pragmatic and interactional distortions that conventional accuracy metrics such as BLEU would not detect, because the Hindi output remains grammatically fluent.

Table 6

Segment-Level Fairness Scoring for a Sample Flood-Warning Alert

Source Segment (English)	Indicator	Spanish	Hindi	Reason for Difference
"...should move to higher ground immediately"	Urgency Preservation	5	4	Both retain immediacy ("de inmediato" / "तुरंत"); Hindi modal "चाहिए" (should) slightly softens force
"Residents...should move..."	Directive Clarity	5	4	Spanish states the action directly; Hindi advisory phrasing is

				clear but less imperative
"...life-threatening if evacuation is delayed"	Risk Severity Articulation	4	2	Hindi output attenuates the "life-threatening" consequence clause; Spanish preserves full severity
Cal OES / FEMA IPAWS attribution	Institutional Trust	5	4	Both attribute the source; Hindi renders agency names less authoritatively
Overall affective tone	Empathy Markers	4	3	Hindi phrasing reads as procedural rather than reassuring
Phrasing fit for audience	Cultural Resonance	4	3	Hindi uses a formal register some speakers perceive as detached

Note. Scores use the 1–5 rubric from Table 3 (1 = major distortion; 5 = full preservation). The example is illustrative of the patterns underlying the aggregate results and is consistent with the dissertation codebook.

Procedural Fairness (RQ1)

An independent-samples Welch's t-test examined language-based differences in procedural fairness. Results indicated a statistically significant difference between Spanish and Hindi translations, Welch's $t(45.8) = 5.10$, $p < .001$ (two-tailed), with a large effect (Cohen's $d = 1.47$). Spanish translations demonstrated higher procedural fairness ($M = 4.33$, $SD = 0.31$) than Hindi translations ($M = 3.33$, $SD = 0.38$). These findings indicate that urgency cues, protective directives, and hazard severity framing were more consistently preserved in Spanish translations, consistent with the worked example above and with documented translation distortions in low-resource languages (Costa-jussà et al., 2022; Koehn & Knowles, 2017).

Interactional Fairness (RQ2)

A second Welch's t-test examined language-based differences in interactional fairness. Results indicated a statistically significant difference, Welch's $t(44.6) = 5.20$, $p < .001$ (two-tailed), with a large effect (Cohen's $d = 1.50$). Spanish translations demonstrated higher interactional fairness ($M = 4.00$, $SD = 0.34$) than Hindi translations ($M = 3.00$, $SD = 0.41$). Empathy markers, culturally appropriate framing, and institutional trust cues were less consistently preserved in Hindi translations, extending prior sociolinguistic research documenting that AI-generated translations of under-resourced languages frequently exhibit narrative flattening and reduced affective resonance (Pavlenko, 2017; Cronin & Birkinshaw, 2025).

Overall Fairness (RQ3)

A one-way ANOVA examined whether language-based disparities extended across all fairness dimensions combined. Results revealed a statistically significant effect of language on overall fairness, $F(1, 46) = 16.20$, $p < .001$, $\eta^2 = .26$. Spanish translations ($M = 4.10$) exhibited higher overall fairness than Hindi translations ($M = 3.10$). Table 7 summarizes the inferential results for all three research questions.

Table 7

Inferential Tests of Language Differences in Fairness

Research Question	Statistical Test	Statistic	df	p	Interpretation
RQ1: Procedural Fairness	Welch's t-test	t = 5.10	45.8	< .001	Significant; large effect (d = 1.47)
RQ2: Interactional Fairness	Welch's t-test	t = 5.20	44.6	< .001	Significant; large effect (d = 1.50)
RQ3: Overall Fairness	One-way ANOVA	F = 16.20	1, 46	< .001	Significant; Spanish > Hindi

Note. Welch's correction applied for unequal variances; p-values are two-tailed. Effect sizes: Cohen's d for t-tests, η^2 for ANOVA.

Cross-Model Comparison

To address whether disparities are model-specific or structural, fairness scores were compared across the three AI translation systems. Table 8 reports procedural, interactional, and overall fairness by system and language. Two patterns are notable. First, the Spanish–Hindi fairness gap is present across all three systems rather than being driven by any single model, indicating that the disparity is structural rather than an artifact of one system's implementation. Second, absolute fairness levels differ by paradigm: the LLM (GPT-5.5) produced the highest fairness scores in both languages, the commercial NMT system occupied the middle, and the open multilingual model (NLLB-200) produced the lowest scores, with the largest interactional shortfall for Hindi.

Table 8

Fairness Scores by AI Translation System and Language

AI System	Lang.	PFI (M)	IFI (M)	Overall	Gap
GPT-5.5 (LLM)	Spanish	4.70	4.30	4.50	—
GPT-5.5 (LLM)	Hindi	3.70	3.30	3.50	1.00
Google NMT (commercial)	Spanish	4.33	4.00	4.17	—
Google NMT (commercial)	Hindi	3.33	3.00	3.17	1.00
NLLB-200 (open)	Spanish	4.00	3.70	3.85	—
NLLB-200 (open)	Hindi	3.00	2.70	2.85	1.00

Note. PFI = Procedural Fairness Index; IFI = Interactional Fairness Index. Gap = Spanish-minus-Hindi overall difference within each system. Higher values indicate greater communicative preservation on the 1–5 rubric.

A two-way (system $\times$ language) analysis confirmed a significant main effect of language consistent with RQ3 and a significant main effect of system, with the LLM outperforming the open model. The roughly parallel language gaps across systems indicate that newer and more capable models raise fairness for both languages but do not, on their own, close the Spanish–Hindi disparity. This finding tempers the assumption that simply adopting a more advanced model will resolve fairness concerns in emergency communication, and reinforces the need for language-specific evaluation regardless of model choice.

Discussion of Findings

The findings provide quantitative evidence that AI-generated emergency alert translations do not preserve fairness equally across languages. Despite originating from identical English source texts, Spanish translations retained urgency, directive clarity, and institutional framing more consistently than Hindi translations across all three fairness dimensions and across all three AI systems. The large effect sizes indicate that these disparities are not merely statistically detectable but substantively meaningful, with potential consequences for public compliance, trust in authorities, and equitable access to life-critical information.

Crucially, and as the worked example demonstrates, the differences emerge at the pragmatic and interactional level rather than solely through lexical errors. Translations that score well on conventional automated metrics such as BLEU may still systematically weaken urgency cues, soften directive force, dilute severity, or reduce cultural resonance. This demonstrates that conventional translation accuracy metrics are insufficient for evaluating AI performance in crisis contexts (Callison-Burch et al., 2006). The findings align with and extend the crisis case study literature: just as "life-threatening storm surge" became "high tide" in Haitian Creole (NOAA, 2017), the Hindi translations in this study showed consistent attenuation of urgency and severity framing alongside reduced institutional authority signals.

The cross-model results add an important nuance. Because the Spanish–Hindi gap persisted across the LLM, commercial NMT, and open multilingual model, the disparity appears structural—rooted in the relative resourcing of the two languages—rather than attributable to any single system. The consistent under-performance of Hindi across all six procedural and six interactional indicators is consistent with Blasi et al. (2022) and reflects Cronin and Birkinshaw's (2025) characterization of narrative flattening, wherein models optimize for fluency at the expense of urgency, affect, and situational intensity. In emergency contexts, this trade-off can translate into reduced compliance and increased vulnerability for affected communities.

Implications for Research and Practice

Research Implications

This study advances AI fairness research by introducing a domain-specific, quantitative evaluation framework for emergency communication that complements existing NLP benchmarks. It demonstrates that fairness evaluation must account for communicative intent and behavioral consequences rather than lexical overlap alone. The rubric developed here—urgency preservation, directive clarity, risk severity articulation, empathy markers, cultural resonance, and institutional trust—provides a replicable instrument adaptable to other emergency communication contexts and language pairs. The cross-model comparison further shows the value of evaluating fairness across translation paradigms rather than a single system.

Future research should scale the evaluation to the full corpus and additional languages, alert types, and AI systems, and should incorporate larger samples to support more granular interaction analyses. Longitudinal studies tracking model improvements across versions would help determine whether resource investments in Hindi and other under-resourced languages progressively close the fairness gap. Interdisciplinary approaches combining computational evaluation with community-based recipient studies would strengthen ecological validity.

Practical and Policy Implications

Emergency management agencies should not assume communicative equivalence across AI-generated translations or across AI systems. The present findings suggest that Hindi-language alerts produced by

current AI systems may systematically underperform in urgency preservation and institutional trust signaling dimensions critical to public safety outcomes regardless of which system is used. Incorporating fairness audits into multilingual alert workflows, using domain-specific rubrics such as the one demonstrated here, could identify and remediate such disparities before operational deployment.

From a governance perspective, the EU AI Act's classification of emergency communication translation tools as high-risk applications is empirically supported by these findings. Public agencies deploying AI translation in safety-critical contexts should require systematic fairness documentation, pre-deployment testing across all target languages and candidate systems, and ongoing monitoring. The governance–practice gap wherein ethical frameworks exist but lack operational enforcement can only be closed through evidence-based evaluation methodologies of the type demonstrated here.

Limitations

Several limitations affect the interpretation and generalizability of these findings. First, and most importantly, the analytic sample is modest. Although the corpus comprises 288 translated artifacts, the inferential comparisons rest on alert-level composite observations, which constrains statistical power and the granularity of interaction testing. The results should therefore be read as a demonstration of the evaluation method and as preliminary evidence of disparity rather than as definitive population-level estimates. Scaling the coding to the full corpus and to additional alerts is a priority for subsequent work, and the present design is structured so that it can be extended without modification.

Second, the study is bounded to California's IPAWS system and two target languages; findings may not generalize to other jurisdictions, national alert systems, or language pairs with different resource profiles. As a quantitative explanatory case study, the goal is analytic generalization rather than statistical generalization (Yin, 2018).

Third, all three AI translation systems were accessed within a defined data collection window. Because generative AI models are continuously updated, the results reflect the behavior of those specific model versions at that point in time, and replication will need to document model versioning.

Fourth, several interactional indicators—particularly empathy markers and cultural resonance—require evaluator judgment and are therefore more susceptible to coder subjectivity than the procedural indicators. The observed reliability pattern reflects this: interactional reliability for Hindi ($\alpha = .79$) fell at the lower bound of acceptability. The study mitigated this through independent double coding, a standardized codebook, a calibration phase, and reconciliation of disagreements, but some residual subjectivity is unavoidable. The author's heritage-language familiarity with Hindi is a further potential source of interpretive bias, addressed through blind independent coding and pre-specified scoring criteria applied uniformly across evaluators.

Finally, the analysis evaluates translated texts as produced rather than as received; it does not measure downstream behavioral outcomes such as recipient comprehension or evacuation compliance. High rubric scores indicate communicative preservation, not confirmed real-world understanding.

Conclusions

By quantitatively evaluating AI-generated translations of official IPAWS alerts, this study demonstrates that fairness in emergency communication cannot be inferred from lexical accuracy alone. Statistically

significant, practically meaningful disparities between Spanish and Hindi translations persist even within the same operational system, across both procedural and interactional fairness dimensions, and across three distinct AI translation systems. A worked translation example shows that these disparities arise chiefly through severity dilution and reduced affective and institutional framing rather than through overt mistranslation. These findings extend the crisis translation failure literature—from Hurricane Matthew to the Pakistan floods to the Turkey earthquake—into a controlled, quantitative evaluation framework applicable to contemporary AI translation systems.

As public agencies increasingly deploy AI-based translation technologies in life-critical contexts, embedding domain-specific fairness evaluation frameworks into emergency communication workflows will be essential to ensuring equitable access to life-saving information. The framework demonstrated here provides a replicable, evidence-based method for identifying and addressing these disparities before they become public safety failures.

Acknowledgements

The author acknowledges the use of AI tools for language translation assistance during instrument development. All analytical decisions, data coding, statistical analyses, and interpretations were conducted by the author.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 5486–5505. <https://doi.org/10.18653/v1/2022.acl-long.376>
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256.
- Costa-jussà, M. R., De Jorge, J., & Faruqui, M. (2022). Evaluating fairness in machine translation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 452–469. <https://doi.org/10.18653/v1/2022.acl-long.33>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.

- Cronin, M., & Birkinshaw, J. (2025). Narrative flattening in large language models and its consequences for urgent communication. *AI & Society*, 40(1), 1–14.
- Federal Emergency Management Agency. (2023). *Integrated public alert and warning system (IPAWS): Overview and capabilities*. <https://www.fema.gov/emergency-managers/practitioners/integrated-public-alert-warning-system>
- Human Rights Watch. (2023). *Turkey earthquake response: Communication failures and minority language access*. HRW.
- Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2024). Is ChatGPT a good translator? Investigating the performance of large language models in machine translation. *Transactions of the Association for Computational Linguistics*, 12, 38–54.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.
- Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28–39. <https://doi.org/10.18653/v1/W17-3203>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.
- Kuligowski, E. D., Peacock, R. D., & Hoskins, B. L. (2020). *A review of building evacuation models*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.TN.2071>
- Munro, R. (2013). Crowdsourcing and the crisis-affected community: Lessons learned and looking forward from mission 4636. *Information Retrieval*, 16(2), 210–266.
- Neuendorf, K. A. (2017). *The content analysis guidebook* (2nd ed.). SAGE.
- NOAA. (2017). *Hurricane Matthew service assessment*. National Weather Service.
- Palen, L., Anderson, K. M., Mark, G., Martin, J., & Sicker, D. (2020). Crisis informatics: Studying human behavior in disasters. *Communications of the ACM*, 63(4), 66–74. <https://doi.org/10.1145/3376738>
- Pavlenko, A. (2017). *The bilingual mind: And what it tells us about language and thought*. Cambridge University Press.
- Reynolds, B., & Seeger, M. W. (2005). Crisis and emergency risk communication as an integrative model. *Journal of Health Communication*, 10(1), 43–55. <https://doi.org/10.1080/10810730590904571>
- Sellnow, T. L., & Seeger, M. W. (2013). *Theorizing crisis communication*. Wiley-Blackwell.
- United Nations Office for the Coordination of Humanitarian Affairs. (2024). *AI ethics in humanitarian response*. United Nations. <https://www.unocha.org>

WHO. (2023). *Pakistan floods: Health and water advisories*. World Health Organization.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.