AI system review:

Using the City's Google AutoML translation system, residents can view the San José 311 website (https://311.sanjoseca.gov) and file service requests in Vietnamese or Spanish. The City will receive the service request in English and respond in English. The response will then be translated into the resident's preferred language.

The translation system is built on Google's base translation model and further trained by City-provided data relevant to City service requests. Consequences for poor performance are limited to loss of services, and inaccurate translations can be fixed by the City manually updating the model.

Additionally, the system is easy to update and monitor for accuracy through the vendor's cloud platform. Based on standard evaluation metrics of translation AI (the BLEU metric), the system performs well in English to Spanish, Spanish to English, and English to Vietnamese. However, the system performs relatively poorly (though still usable) when translating from Vietnamese to English. The City should explore improving its Vietnamese to English translation database for better translations.

Given the easy access to update the system, addressable set of consequences, and understandable accuracy metrics, this AI system is approved for usage in the City. The City should work to improve its Vietnamese to English translation.

Vendor FactSheet for Algorithmic Systems

Please provide details regarding your algorithmic system product by filling out the FactSheet[1] template and Algorithmic Impact Assessment Questionnaire below.

FactSheet

Vendor Name	Google
Model Name	Auto ML Translation
Overview	The system is based on the standard Auto ML Google translation model. City staff can customize the system to their specific needs by training the system on sentence pairs in English, Vietnamese, and Spanish. The tool is used to translate customer messages in the San José 311 non-emergency helpline (SJ311) service. SJ311 can be accessed via phone, online,[2] and through the SJ311 mobile application.
Purpose	The system is used to translate customer messages to and from English, Vietnamese, and Spanish in the SJ311 service's chat function.
Intended Domain	Natural language processing
Training Data	The base model is trained on millions of examples of sentence pairs for 133 languages. The training data for the customized SJ311 model are sentence pairs for each language combination (English-Vietnamese, English-Spanish). In addition to basic language, the sentences feature vocabulary that is highly relevant for common SJ311 reporting areas (abandoned vehicles, illegal dumping, potholes, etc.).
Model Information	Auto ML Translation enables clients to perform supervised learning, which involves training a computer to recognize patterns from translated sentence pairs. [3] Using supervised learning, clients can train a custom model to translate domain-specific content they care about (i.e., San Jose city services).[4]

[1] The FactSheet template is heavily inspired by the IBM Research AI FactSheets 360 project.
[2] San José 311 website: https://311.sanjoseca.gov/
[3] https://cloud.google.com/translate/automl/docs
[4] https://cloud.google.com/translate/automl/docs/beginners-guide

Inputs and Outputs	**Input:** A text sentence in English, Vietnamese, or Spanish. **Output:** Depending on the language translation needed, a text sentence in English, Vietnamese, or Spanish.
Performance Metrics	BLEU[5] score (Vietnamese to English): 34.13 BLEU score (English to Vietnamese): 74.37 BLEU score (Spanish to English): 67.38 BLEU score (English to Spanish): 57.7 A rough guideline of BLEU score interpretation is below:

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Optimal Conditions	No specific documentation provided. We can infer that the model will perform optimally in conditions similar to the training data it was provided. In this case, the model is the standard google translate model with additional training data on communications relevant to San José 311,[6] or City services.
Poor Conditions	No specific documentation required. We can infer that the model will perform poorly in conditions not similar to the training data it was provided. In this case, the model is the standard google translate model with additional training data on communications relevant to San José 311,[7] or City services. For example, the model may not perform well in translating conversations pertaining to legal matters unrelated to City services.
Bias	There appear to be a varying quality of performance across languages and the direction of translation (i.e., Vietnamese to English vs. English to Vietnamese).[8]

[5] How to interpret BLEU scores
[6] More information on San José 311 can be found at https://311.sanjoseca.gov/
[7] More information on San José 311 can be found at https://311.sanjoseca.gov/
[8] According to Google, trying to compare BLEU scores across different corpora and languages is strongly discouraged. Even comparing BLEU scores for the same corpus but with different numbers of reference translations can be highly misleading.

	The out-of-the-box Google Translate service has been shown to suffer from gender bias by changing the gender of translations when they do not fit with common stereotypes.[9] Google has an Inclusive ML guide for clients as they use Auto ML systems for their customized applications that includes best practices to promote fairness in ML outcomes.[10]
Test Data	The testing data are sentence pairs for each language combination (English-Vietnamese, English-Spanish). In addition to basic language, the sentences feature vocabulary that is highly relevant for common SJ311 reporting areas (abandoned vehicle, illegal dumping, potholes, etc.).

Algorithmic Impact Assessment Questionnaire

Accuracy	
Under what conditions/circumstances has the system been tested?	• Testing data includes common vocabulary and vocabulary particularly relevant to typical SJ311 inquiries.
Have the vendors or an independent party conducted and published a validation report (including the methodology and results) that audits for accuracy and discriminatory/disparate impact? If yes, can the City review the study?	• There are compliance reports from third parties, but no specific report published on the translation models.
Will the model be learning from the information it gets in the field during deployment?	• Model can be trained by users with additional training data to customize for specific application. For example, San José adds data to improve performance on translations specific to City services. • The base model is regularly updated by vendor.
Equity	
What quality control is in place to test and monitor for potential biases in the AI system (e.g., non-representative training data, overfitting, hard-coded rules)?	• Google's Ethical AI team reviews new AI tools before they launch, with an eye towards possible bias that may be present in the system.[11] Reviewers may recommend technical evaluations (e.g.,

[9] https://algorithmwatch.org/en/google-translate-gender-bias/
[10] https://cloud.google.com/inclusive-ml
[11] https://ai.google/responsibilities/review-process/

	checking for unfair bias in ML models). If necessary, they consult with trusted external advisors (e.g., human rights experts). Adjustments are made after reviewers offer mitigation strategies.
How can the City and its partners flag issues related to bias, discrimination, or poor performance of the AI system?	• City can open a support case and Google Support will take the feedback into account.
Explainability	
What performance metrics were selected to judge the model's effectiveness? What is it optimizing for, and under what constraints?	• BLEU metric (specific to Translation) • Google's confidence score (relevant to any Auto ML prediction)
How are the outcomes of the AI system explained to subject matter experts, users, impacted individuals, or others?	• Each translated sentence is presented with a confidence score from 0-100% that shows the model's "confidence" in that translation. While the number might not be a perfect interpretation of the BLEU score, it provides individuals with the gist of how good a translation likely was.