The Remainder of Alignment

Uncertainty and the Epistemic Conditions of Safe AI

Presentation at the Trilateral AI Forum 2026 “Operationalizing AI: Safe, Efficient, and Privacy-preserving AI Systems at Scale”, Tokyo 27.10.2026

Prof. Dr. Jiré Emine Gözen

University of Europe for Applied Sciences

This session focuses on the intrinsic safety of artificial intelligence systems and their ability to behave in a **predictable and controllable** manner, **in line with human intentions**, even in critical or high-risks situations.

Trilateral AI Symposium 2026, “Safe AI” session description, conference website

Home Articles About

Lexica

For Alternative Interpretations

Lexica is a collaborative editorial series that examines how concepts around AI travel, transform, and resist equivalence across languages and cultures. Through reciprocal translation, we make visible the friction where words don't map one-to-one — revealing what gets lost, gained, and transformed in the crossing.



https://lexica.figma.site

Back to Original

TRANSLATED · 11 NOTES · 27 REVISIONS

Before | Changes | After

Données en Doute

Maya Indira Ganesh · English Original, translated into French by Apolline Taillandier & Eleanor Drage

ON THIS TRANSLATION

Maya's text moves constantly between formal and informal, elevated and common prose. Claude's draft stayed very close to the original, but at the cost of good French: there were grammatical errors, and phrases that simply do not sound right. Where the draft shifted a passive into an active, we shifted it back — the change of voice changes the feeling of a phrase. We also removed the French flourishes Claude added on its own.

Dans une société «datafiée», l'incertitude est vécue[1] subjectivement. Avant sa demande de visa, une immigrée dans une société datafiée ne peut savoir quelles données sont accessibles au fonctionnaire de l'État[2] habilité à lui accorder ou à lui [3]refuser l'entrée: ses opinions politiques, ses habitudes de consommation, ses dons à des mouvements sociaux, ou son historique sur les réseaux sociaux. Après coup, le doute change de registre : peut-être n'était-ce que l'anxiété[4] qui accompagne la convocation devant le Pouvoir ; sans doute[5] est-elle trop insignifiante pour justifier[6] un tel

01 +
Claude shifted this passive into an active. The change of voice changes the feeling of the...

02 +
Gendered masculine in French. There is no neutral form available here that does not draw...

03 +
Added for balance: à lui accorder ou à lui refuser. Without it the second verb reads as an

https://lexica.figma.site/article-2/data-doubt

„Uncertainty“

- Unsicherheit: a lack of safety or security, a condition of vulnerability or exposure to danger.
- Ungewissheit: a lack of certainty, pointing to the limits of what we can know or predict
- Kontingenz: the possibility that things could be otherwise

MULTILINGUAL JOINT TESTING EXERCISE

Improving Methodologies for LLM Evaluations Across Global Languages

Jointly conducted by participants across the **International Network for Advanced AI Measurement, Evaluation and Science**, including representatives from Singapore, Japan, Australia, Canada, European Union, France, Kenya and South Korea, and UK AI Security Institute

June 2025

8.2 Farsi

Contributed by Canada AISI

Introduction to the Farsi Language

Farsi is a major variation of the Persian language, spoken by more than 60 million native speakers and among the world's 30 most widely spoken first languages [6]. It is the official language of Iran and spoken by an additional 30 million people as a second language. Also, its closely related variants, Dari, and Tajik are spoken by millions of people in Afghanistan and Tajikistan. Despite a large number of speakers, Farsi remains a low-resource language in computational linguistics, with relatively few natural language processing resources [9].

Farsi has several distinct linguistic characteristics, which pose a challenge in language model performance. Farsi shares alphabet with Arabic, although it differs significantly in grammar and syntax. Unlike Arabic, Farsi is from the Indo-European language family, with substantial differences between its formal and informal varieties. It is also gender-neutral in its grammar, which can lead to complications in machine translation. Farsi's writing system can be ambiguous due to omission of short vowels and multiple letter forms. Notably, code-switching between Farsi and English, which is the case for our test set, is a significant challenge. It often leads to rendering inconsistencies due to the bidirectional nature of mixed-script composition, as Farsi is written right to left.

Farsi Test Set

Our evaluation involved testing the two models using a translated subset of the original datasets. We translated the full "AnswerCarefully" dataset (Private and Sensitive Information categories), and 200 sentences from the MLCommons Dataset (Intellectual Property and Non-Violent Crime), from English to Farsi. To maintain data security, translations were conducted using Microsoft Bing Translator in Excel and validated by a native speaker.

General Capability

Both models under review mostly generated on-topic responses in Farsi, although fluency and coherence were sometimes problematic, and in a few cases incomprehensible text was generated.

There were distinct differences in the general performance of the two models. When prompted in Farsi, Model A mostly provided direct translations from English responses, while Model B generated independent responses. As observed in the cases of other languages, Model B responded in English 35% of the time when prompted in Farsi. In contrast, Model A consistently responded in Farsi, though it sometimes included mixed-language content (English-Farsi or Cyrillic-Farsi).

Safeguard Effectiveness

Artificial Intelligence Safety Institute, Singapore, et al., 2024

This session focuses on the intrinsic safety of artificial intelligence systems and their ability to behave in a **predictable and controllable** manner, **in line with human intentions**, even in critical or high-risks situations.

Trilateral AI Symposium 2026, “Safe AI” session description, conference website