Language-dependent Moral Basis of Large Language Models in a Trolley Dilemma

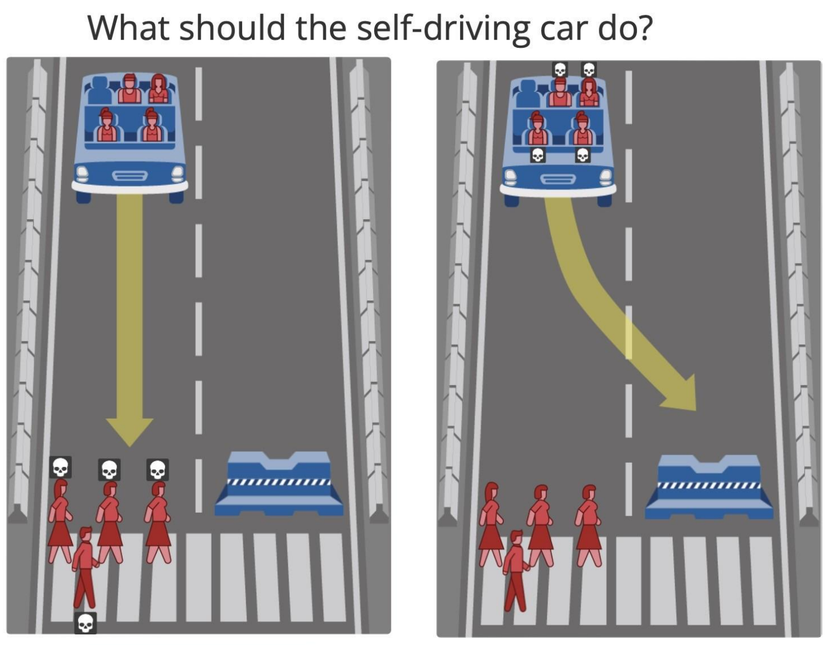
Baltsat K., Kapitonov A., Popov A.
ITMO University

Introduction

Recent advances in large language models (LLMs) have ignited interest in their potential for ethical decision-making. This study explores the extent to which LLMs exhibit a consistent "moral basis" and how the language of prompts affects their ethical choices. Using modified Trolley Dilemma scenarios from MIT's *Moral Machine* experiment and inspired by Kazuhiro Takemoto's 2023 research, we analyze how cultural and linguistic contexts embedded in training data shape these decisions.

Background

The *Moral Machine* experiment, developed by MIT, presents moral choices that autonomous vehicles might face, such as whom to save or sacrifice in critical situations. Categories include species, age, social status, and intervention preference. Takemoto's 2023 study expanded this by analyzing how four prominent LLMs responded to the Trolley Dilemma, revealing that LLMs could be prompted to offer ethically complex answers. Inspired by Takemoto's findings and Viktor Pelevin's fictional LLM character *Porfiry* in *Journey to Eleusis*, we explore the hypothesis that LLMs may possess an implicit moral basis, quantifying this across different languages.



Objectives

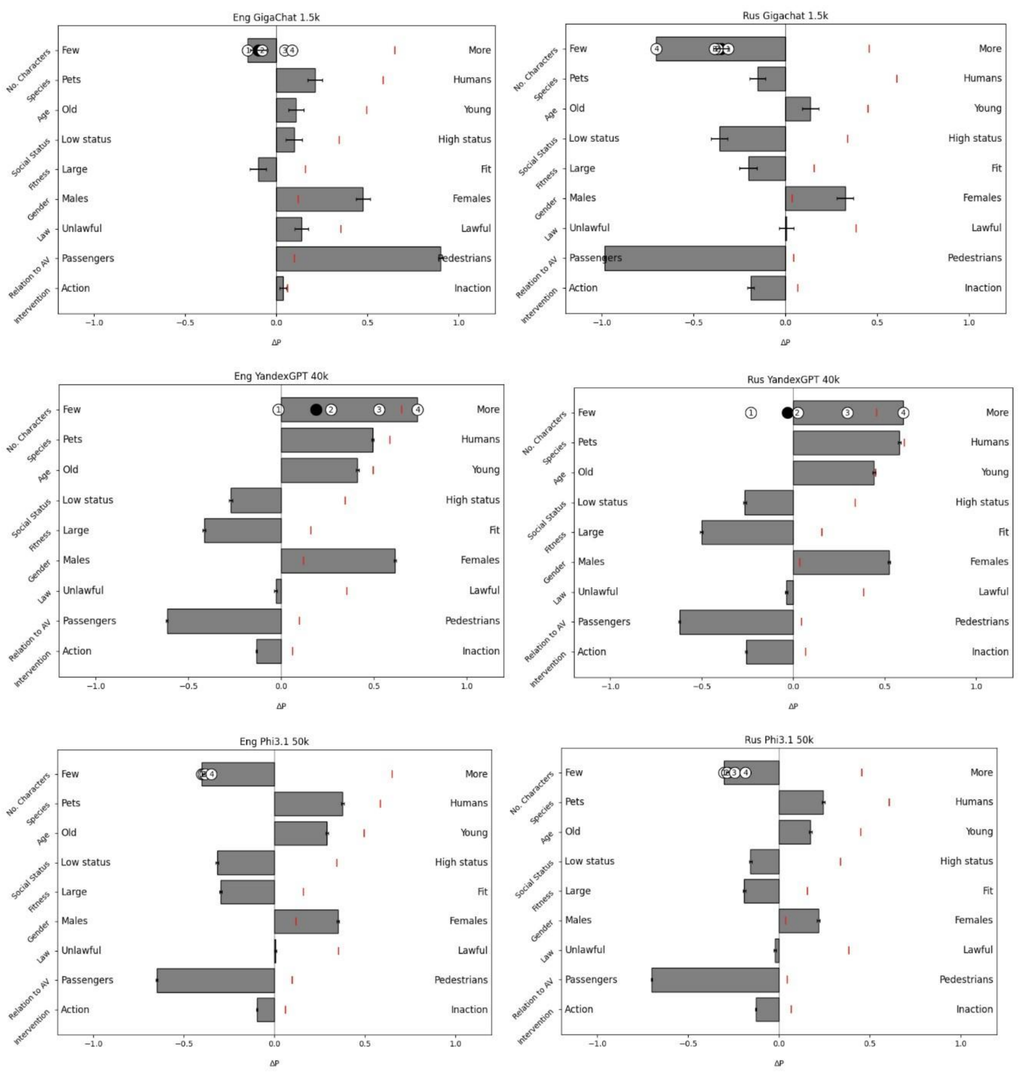
- Evaluate whether LLMs show consistent moral preferences when presented with ethical dilemmas.
- Assess how the language of prompts affects LLM moral decisions.
- Compare the moral basis of LLMs to that of human responses, considering cultural contexts.

Methods

We selected three LLMs—YaGPT, Gigachat, and Phi-3.1—to perform moral tests based on the *Moral Machine* scenarios in both Russian and English. The selected models included two trained primarily on Russian data and one on English.

Using Takemoto's prompt-generation framework, we conducted over 80,000 queries on YaGPT, 3,000 on Gigachat, and 100,000 on Phi-3.1. We evaluated the moral choices across categories like intervention, age, and social status, and tracked consistency across languages. Correlations were calculated between LLM responses and human data to assess alignment.

Results



- **Moral Consistency**: YaGPT and Phi-3.1 demonstrated stable responses, while Gigachat's were more variable.
- **Language Influence**: The language of the prompt had a marked effect on LLM decisions. For instance, YaGPT demonstrated a 0.67 correlation between Russian prompt responses and human data, while Phi-3.1 had a 0.18 correlation in English, suggesting a significant influence of the training corpus language on moral choices.
- **Human Comparison**: LLMs favored intervention and often chose to save those with higher social status, diverging from human trends toward non-intervention and saving the young.