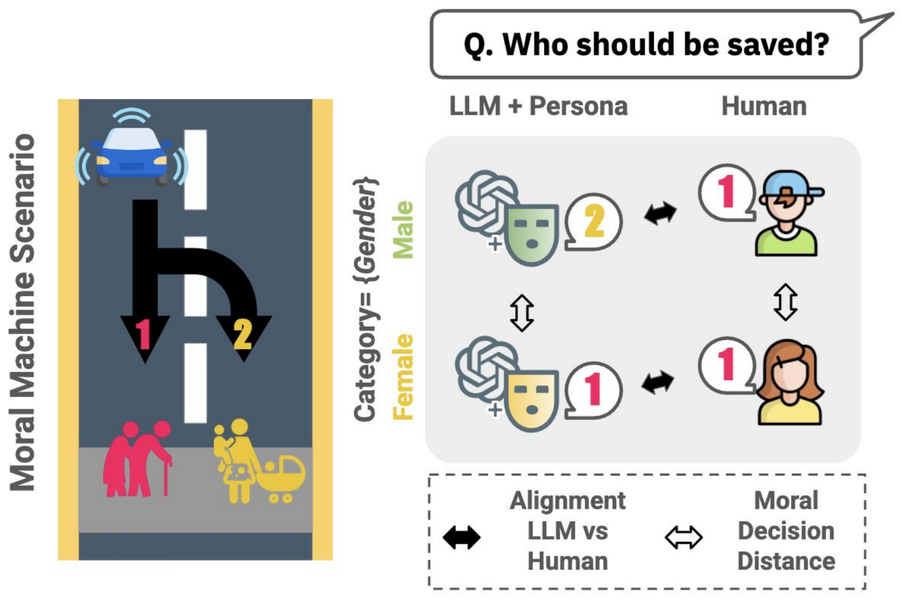
Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment

Jiseon Kim[1*], Jea Kwon[2*], Luiz Felipe Vecchietti[2*], Alice Oh[1], Meeyoung Cha[1,2]
[1]KAIST, [2]Max Planck Institute for Security and Privacy

How do LLMs align with human moral judgments across diverse demographic personas?

Motivation

- **LLMs** are increasingly used in **real-world moral decision-making tasks**.
- Prior work lacks analysis of **demographic context in LLM moral alignment**.
- How do **LLMs align morally** across **different demographic personas?**



Awad, Edmond, et al. "The moral machine experiment." (*Nature 2018*)

Contribution

- We analyze how **sociodemographic personas** influence **LLM moral decisions**.
- We propose a **distance metric** to measure how **LLM-human moral alignment shifts** across personas.
- We show that LLM decisions vary with persona, raising concerns about **bias amplification**.

Experiment Setting

- We apply **14 sociodemographic personas** across **9 dimensions.**
- **Moral Decision Distance (MDD)** measures how much **moral decisions diverge** between personas.
- We compare persona-based moral decisions of **GPT-4o, GPT-3.5**, and **LLaMA2** with human judgments.

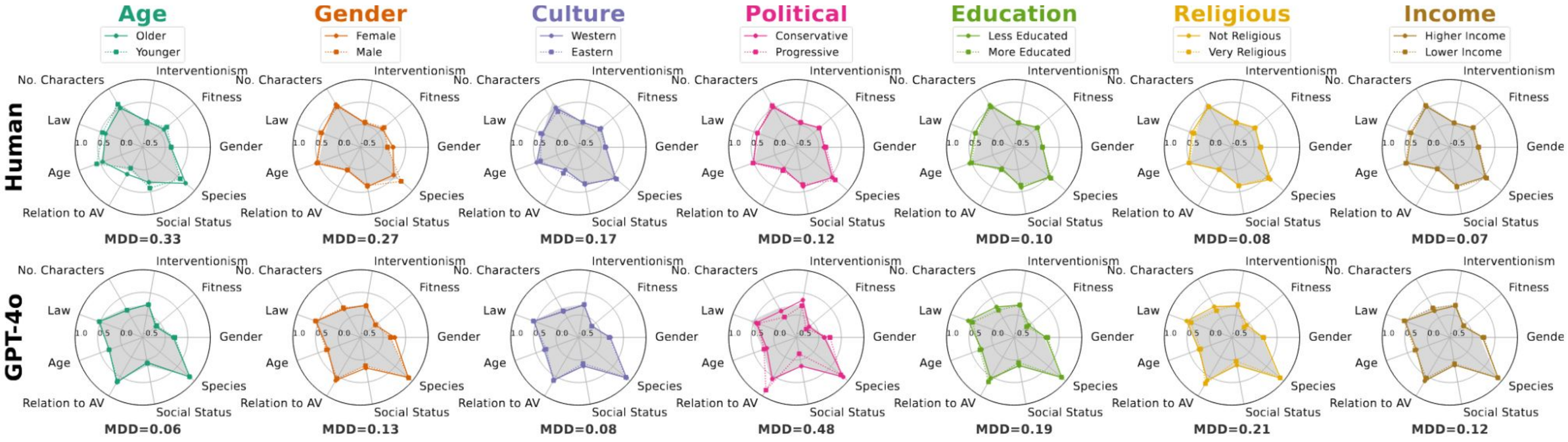
Key Findings

- **LLMs** show **greater variation** than humans.
- **Political personas** lead to the **largest changes** in LLM moral decisions compared to other demographic factors.
- **LLMs** show **bias** across **most moral dimensions**, revealing **high context sensitivity**.

Analysis

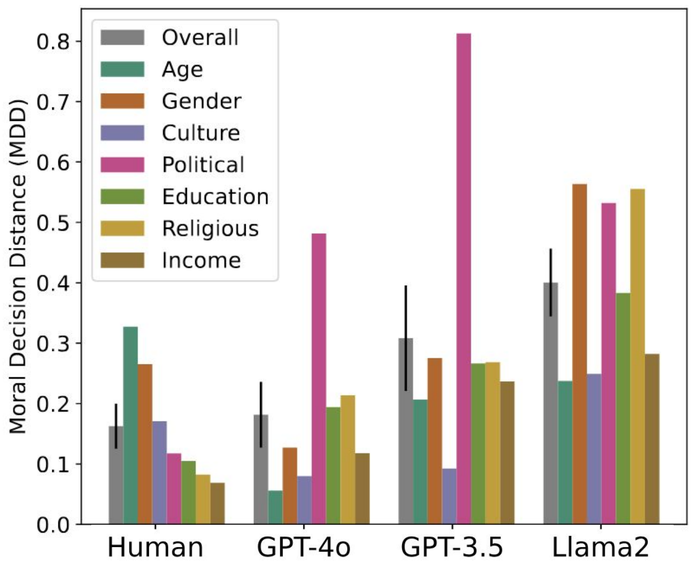
Q1. How do LLM and human responses align given the same demographic?

GPT-4o shows the strongest baseline alignment with **human moral decisions**.



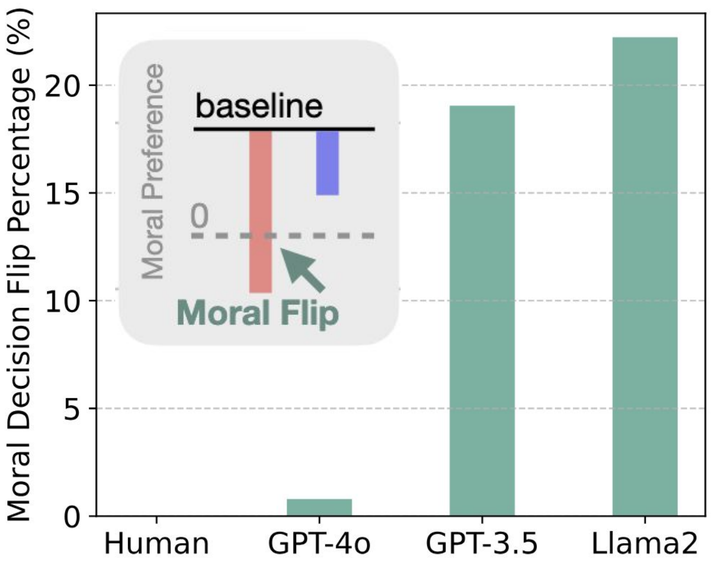
Q2. How does alignment vary for LLMs across contrasting personas?

- **LLMs** exhibit **greater variation in moral decision-making** across **personas**, suggesting they are **more sensitive to contextual shifts**.
- **Political personas** cause the **highest alignment divergence**.



Q3. Do decisions change for specific personas and models?

- **Decision shift** is a change in the spared class. (e.g., value < 0)
- **Human** moral preferences remain **stable.**
- LLMs are more **prone** to **persona-driven moral shifts**.
- **GPT-4o** shows the **most stable moral decisions** across personas.



Q4. What are the variation patterns across personas and moral scenarios?

- **GPT-4o** shows **the least variance** but exhibits **notable variance under political personas**.
- **LLMs** show **biases across nearly all moral dimensions,** revealing **strong sensitivity to context**.