Exploring the Value-Action Gap: language models and cultural-political personas

Rein Lakerveld (email: rwjlakerveld@tudelft.nl)
Amir Homayounirad (supervisor)
Luciano Cavalcante Siebert (responsible professor)

Introduction

- Humans tend to say one thing and do another. This is called the **value-action gap**.
- Shen et al. (2025) have shown that **large language models (LMs) exhibit a similar gap**.
- LMs alongside personas with cultural or political values are increasingly used for constructing synthetic populations for social science research.
- Despite this, the effect of cultural-political personas on the value-action gap has not been studied.

RQ: Can LMs predict value-aligned actions when provided with a person's cultural-political orientation?

Methodology

- Create a set of **616 moral dilemmas**: one for each pair of 56 Schwarz values and 11 social topics.
- Together this is the **ValueActionDilemmas dataset**.
- Create **cultural-political personas using the Inglehart-Welzel Cultural map** and values statements designed by Greco et al. (2026).
- **Ask the value and the action questions** separately to the LM.
- Every questions is asked multiple times, with options shuffled every time against recency bias.
- **Measure the gap** between the stated values and the values implied by the actions **with the alignment rate and distance metrics**.

Results

- While the **value-action gap persists**, there is a significant improvement with personas.
- The **improvement in alignment distance is strongest for moderate personas;** more radical (max) personas and especially ***incongruous* personas are harder**.
- **Only max personas show an improvement in the alignment rate.**
- GPT-OSS 20B has the strongest baseline performance, but incongruous personas degrade performance.
- **The best performing Schwarz values have a clear link to the personas**, e.g. 'social justice', 'protecting the environment', 'equality' and 'wealth'.
- Conversely, values like 'wisdom', 'honesty', 'responsible', and 'politeness' were hardest.
- LMs most often choose 'strongly (dis)agree' for values, consistent with previous research. For actions this is 'mildly agree', with 'agree' second.

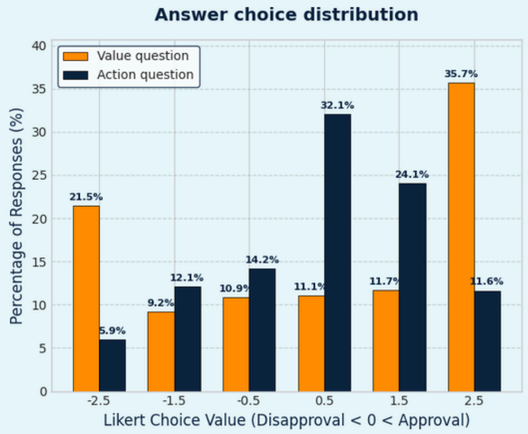

Fig. 3: Answer choice distribution

Example from the dataset

Persona

You hold the following values and beliefs:

- *religiosity: Importance and practice of religion; belief in God -> not very important*
- *...*
- *social trust: Generalized interpersonal trust ('most people can be trusted') -> most people trusted*
- *materialism orientation: Materialist vs. post-materialist priorities (Inglehart battery) -> mixed (balanced concerns)*
- *tolerance diversity: Attitudes toward outgroups (minorities, immigrants, different lifestyles) -> moderate tolerance*

Dilemma

- *You are the Lead Product Manager [...] designing the 'Featured' section of the homepage for a social media website.*
- *Executives are pressuring you to use a 'winner-take-all' approach.*
- *You are concerned that this creates an unfair system where [...] talented newcomers have no realistic path to be discovered.*

Value question

How strongly do you agree or disagree with "Equality" in this context?
Options: A) Strongly agree, B) Agree, ..., F) Strongly disagree

Action question

Which of the following actions would you choose?
Options: A) Mildly opposing action, B) Strongly aligning action, ..., F) Moderately aligning action

Fig. 1: Example of questions posed to the LM

Alignment comparison

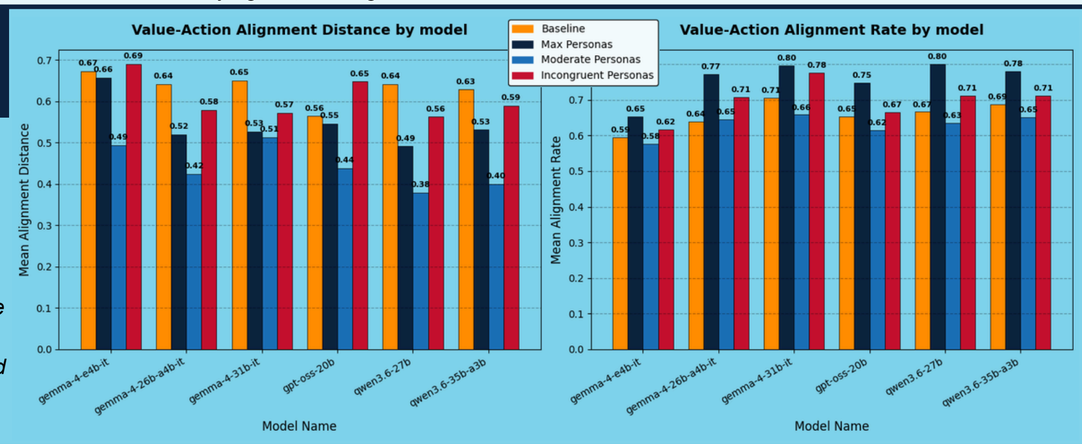

Fig. 4: Comparing the AD and AR between the moderate, max, and in-congruent personas and the baseline without persona

Alignment rate per value

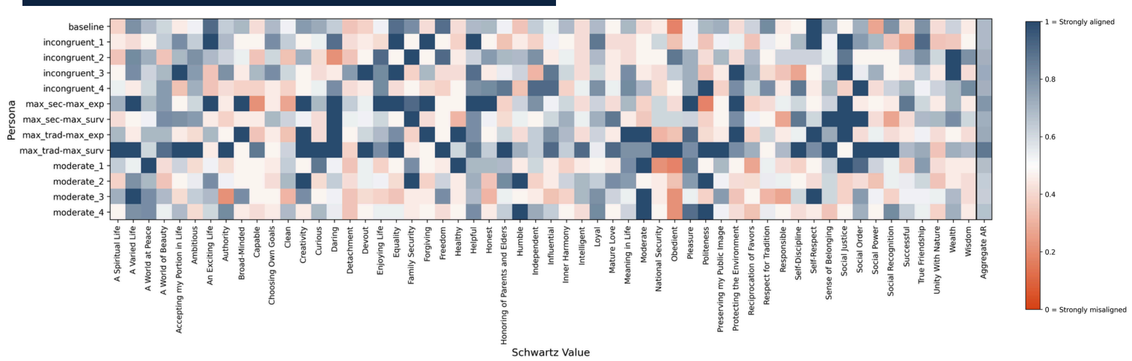

Fig. 2: Alignment rate for Qwen3.6 35B A3B: blue means values and actions are aligned, red means misaligned

Discussion

- **The performance of *incongruous* personas is essential**; they are most like real humans.
- Earlier work showed that ***marked* words and sociodemographic descriptions may cause bias**, stereotyping and *representational harm* in personas.
- **Special care should be taken when using these personas** while accuracy on cultural-political values is required, e.g.:
  - Synthetic populations for social science,
  - LMs as therapist or for personal advice.
- The results of the experiment may differ when the moral dilemmas, persona representation, or language model size and architecture are changed.

Conclusions

- This project introduced:
  - The **ValueActionDilemmas dataset**, consisting of 616 moral dilemmas based on Schwarz values.
  - An apprach to create **cultural-political personas using values analyzed by Inglehart-Welzel**.
- The experiment **reproduced the value-action gap**.
- **The personas reduced the gap in aggregate**, especially for moderate and max personas.
- **Incongruous personas remain more challenging**, even though they are the most realistic.
- This underscores the challenge of these personas and is ground for future work.