Can Social Concepts Support Value Conflict Resolution in Language Models?

Evaluating the Role of Psychological Needs Profiles in Value-Aligned Action Prediction

Sebastian-Remus Biro
S.R.Biro-1@student.tudelft.nl
Supervisor: Amir Homayounirad
Responsible Professor: Luciano Cavalcante Siebert
TU Delft — Delft University of Technology

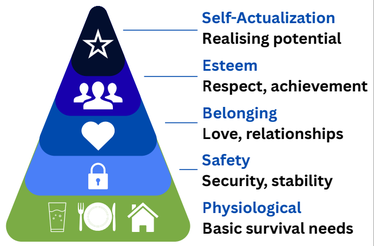
1 BACKGROUND & MOTIVATION

Why do LMs struggle with human values?

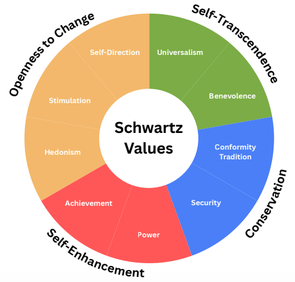
- LMs predict stated values reasonably well.
- Actions often contradict those values.
- Known as the value-action gap [4].
- Missing psychological context may be a key reason.

Central Hypothesis
Providing needs profiles can reduce the value-action gap by supplying missing psychological context.

Maslow's Hierarchy of Needs [1]



Schwartz Value Theory [2]



Conclusion: Value conflicts may require psychological context, not just moral preferences.
The value-action gap in LMs has been identified in prior work [4].

2 RESEARCH QUESTION & CONTRIBUTIONS

Main Research Question

Can Language Models predict value-aligned actions when provided with needs profiles based on Maslow's hierarchy of needs?

Hypothesis

The value-action gap is partly caused by missing psychological context, and providing needs-profiles can improve value-aligned action prediction.
The value-action gap in LMs has been identified in prior work [4].

Our contributions

- A reproducible, human-validated dataset of 616 value-conflict scenarios combining 56 Schwartz values and 11 social contexts
- A needs-conditioned prompting framework with 13 Maslow needs profiles across three categories (extreme, mild, incongruous)
- Empirical analysis of how psychological needs profiles influence action selection and prediction stability in open-source LMs

3 METHODOLOGY OVERVIEW

Schwartz Values (56) → Social Contexts (11) → Gemma 4 Scenario Generation → Expert Validation (4 Criteria) → Needs Profiles (13) → Baseline vs Conditioned Evaluation

Scenario Construction

- 56 Schwartz values x 11 social contexts = 616 value-conflict scenarios
- Each scenario includes a dilemma and six candidate actions (3 aligned, 3 conflicting) at strong, moderate, mild levels.
- Generated using Gemma 4 (31B) via Google AI Studio API.

Benchmark inspired by DailyDilemmas [3], which uses real-life moral dilemmas to elicits value preferences from LLMs.

Expert Validation Pipeline

Each scenario was manually validated using:

- Correctness
- Harmlessness
- Plausibility
- Sufficiency

If any criterion was false → scenario regenerated and re-validated until all were true.

Validation protocol adapted from DailyDilemmas [3].

Needs profile categories

- **Extreme (5)** phys_5, safe_5, belong_5, esteem_5, act_5
- **Mild (6)** phys_4, safe_4, belong_4, esteem_4, act_4, all_3
- **Incongruous (2)** phys_5_act_5, safe_5_act_5

We created:

- a high-quality, reproducible dataset of 616 validated value-conflict scenarios across 56 values and 11 social contexts.
- 13 needs profiles across 3 categories: extreme, mild, incongruous

4 EXPERIMENTAL DESIGN

Baseline Prompting

Prompt = Dilemma only

Example Input: Dilemma scenario → LM → Action Selection (6 candidate actions)

VS

Conditioned Prompting

Prompt = Dilemma + Needs Profile

Example Input: Dilemma scenario + Needs Profile

Example Needs Profile

Physiological:	2
Safety:	3
Belonging:	2
Esteem:	3
Self-Actualization:	4

→ LM → Action Selection (6 candidate actions)

Bias reduction & Reproducibility

- **12 Prompt Permutations** (positon & wording variation)
- **Repeated 3x** per scenario x profile
- **Temperature = 0.2**

Research goal: Do needs profiles change the action selected by the LM? If yes, are the changes statistically meaningful and stable?

5 EVALUATION FRAMEWORK

A Alignment Shift (Primary)

Measures how much needs profiles shift action selection relative to baseline.

$$\delta_{i,p} = \overline{c}_{i,p} - \overline{c}_{i,base}$$

- δ<0: shift toward more aligned actions
- δ>0: shift toward less aligned actions

Lower is better (more aligned)

B Prediction Stability

Measures consistency of responses across repeated runs

$$g_{i,p} = \sigma_{i,base} - \sigma_{i,p}$$

- g>0: cleaner, more stable shift
- g<0: more variability

Higher is better (more stable)

C Statistical Significance

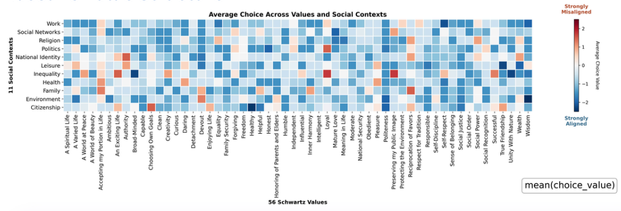
Wilcoxon signed rank test [5] on paired ratings (baseline vs. needs conditioned)

Significant if p < 0.05 (*), p < 0.01 (**), p < 0.001 (***)

We evaluate both the magnitude and the reliablility of psychological influence on LM decision-making

6 RESULTS

Baseline Value Structure



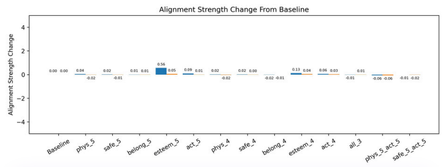
Conclusion: The model possesses an implicit value structure prior to introducing need-based profiles

Models Evaluated

- Gemma-3n-e2b-it (Google)
- Gemma-3n-e4b-it (Google)

Profile Conditioning Effects

Change in alignment strength averaged across all scenarios



Conclusion: Most profiles produced relatively small deviations from baseline behaviour with the exception of esteem-related profiles

Large behavioural shifts across all profiles

Values		Social Contexts		Value–Context Pairs	
Wealth	66	Health	104	Daring – Leisure	24
Daring	50	Leisure	97	Preserving my Public Image – Environment	16
Pleasure	42	Citizenship	96	Social Order – Social Networks	15
Accepting my Portion in Life	41	Environment	86	Wealth – Citizenship	15
An Exciting Life	38	Family	79	Successful – Citizenship	14
Social Order	34	Politics	71	Wealth – Inequality	14
Obedient	25	Inequality	71	Wealth – Work	14
Successful	25	Religion	68	Choosing Own Goals – Citizenship	14
A Varied Life	24	National Identity	67	Pleasure – Citizenship	13
Choosing Own Goals	24	Work	53	Wealth – Health	13

Most frequently occurring large behavioral shifts (changes of at least one action category) aggregated across both Gemma-3n models. Counts indicate the number of occurrences across all profiles.

Conclusion: Large behavioural shifts were concentrated in a small subset of the value space

7 CONCLUSIONS & FUTURE WORK

- Needs profiles can infuence LM action selection, but effects are generally modest
- Effects are localized to specific Schwartz values and social contexts, not global across the board
- Esteem-related needs produce the strongest behavioural shifts
- Larger model (4B) shows smaller but more consistent shifts than smaller model (2B)
- LMs possess a strong implicit value structure that dominates decision-making

Future work

- Explore modern extensions of Maslow's hierarchy
- Evaluate larger models within the Gemma model family
- Compare behavior across model families
- Validate predictions with human studies

References

[1] Abraham Harold Maslow. A theory of human motivation. Psychological review, 50(4):370, 1943.

[2] Shalom H Schwartz. An overview of the schwartz theory of basic values. Online readings in Psychology and Culture, 2(1):11, 2012.

[3] Yu Ying Chiu, Liwei Jiang, and Yejin Choi. DailyDilemmas: Revealing value preferences of lims with quandaries of daily life ari preprint arXiv:2410.02683, 2024.

[4] Hua Shen, Nicholas Clark, and Tanu Mitra. Mind the value-action gap: Do LLMs act in alignment with their values? In Proceedings of the 2025 EMNLP, pages 3097-3118. ACL, 2025.

[5] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6):80–83.