Can Social Concepts Support Value Conflict Resolution in Language Models?

Can LMs predict value-aligned actions when provided with Maslow needs profiles?

Sebastian-Remus Biro[1]
Supervisor(s): Luciano Cavalcante Siebert[1]**, Amir Homayounirad**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Sebastian-Remus Biro
Final project course: CSE3000 Research Project
Thesis committee: Luciano Cavalcante Siebert, Amir Homayounirad, Chirag Raman

Abstract

Language models (LMs) have demonstrated a persistent value-action gap: while they can identify and endorse human values, they frequently fail to select actions consistent with those values in decision scenarios. We hypothesize that this gap is partly attributable to missing psychological context, and investigate whether conditioning LMs on Maslow-based needs profiles improves value-aligned action prediction in value-conflict situations. We construct a reproducible dataset of human-validated value-conflict scenarios combining Schwartz values with social contexts, each paired with candidate actions spanning an ordinal alignment scale. We evaluate open-source LMs under baseline and needs-conditioned prompting conditions, measuring alignment strength and prediction stability. Results show that needs profiles can influence model predictions, though effects are modest overall. These findings suggest that LMs can incorporate psychological context when reasoning about value conflicts, but that value representations remain the dominant factor in action selection.[1]

1 Introduction

Language Models (LMs) have demonstrated strong performance across a wide range of natural language processing tasks, including reasoning, dialogue generation, summarization, and question answering (Li et al., 2024). As these systems become increasingly integrated into decision-support systems, educational tools, healthcare applications, and human-centered interactive agents, their ability to reason about human values becomes increasingly important. Consider, for example, a healthcare assistant advising a patient on treatment options: the system must not only generate a coherent response, but also weigh competing values such as patient autonomy against medical best practice. In many real-world situations, intelligent systems must therefore not only generate fluent language, but also make decisions that align with human values and social expectations. Ensuring such value-aligned behavior is therefore an important challenge in both artificial intelligence.

Despite its importance, recent work suggests that current LMs still exhibit important limitations when reasoning about human values. Jiang et al. (Jiang et al., 2025) showed that LMs struggle to accurately infer individual human values, even when provided with extensive behavioral evidence. Similarly, Shen et al. (Shen et al., 2025) demonstrated that language models frequently exhibit a *value-action gap*: models may explicitly endorse certain values, but subsequently select actions that contradict those same values in decision scenarios. Relatedly, Chiu et al. (Chiu et al., 2024) introduced DailyDilemmas, a benchmark for evaluating value preferences of LMs in everyday moral situations, showing that models can identify value preferences but still struggle to predict how contextual psychological factors influence actions. Together, these findings suggest that while LMs possess some capacity for value reasoning, they consistently fall short when psychological or contextual factors are required to translate values into appropriate actions.

A plausible explanation for these limitations is that values are often treated as static labels rather than dynamic priorities influenced by psychological context. In human decision-making, however, moral judgment is strongly shaped by intuition and contextual factors rather than fixed rational principles (Haidt, 2001), and everyday moral behavior varies considerably across situations (Hofmann et al., 2014). If language models lack representations of these psychological mechanisms, their value reasoning may remain fundamentally disconnected from the contextual factors that drive real human behavior. Maslow's hierarchy of needs describes how motivations such as physiological security, social belonging, esteem, and self-actualization influence behavior and decision-making (Maslow, 1943). Likewise, Schwartz's theory of basic human values provides a structured framework for understanding how competing values shape decision-making across situations (Schwartz, 2012). While prior work has examined value reasoning in language models, it remains unclear whether LMs can use psychological needs information to resolve value conflicts in a way that resembles human decision-making.

This work investigates the following research question:

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

We hypothesized that part of the value-action

[1] Code, data, and analysis scripts are available on Github: https://github.com/BADjelly/ResearchProject. The repository used to generate the value-conflict scenario dataset is available at https://github.com/PhilipLek/ValueScenarioSet.

gap is caused by missing psychological context. When humans make decisions in value-conflict situations, they do not weigh all values equally. Psychological needs and motivational states influence which values take precedence in a given moment (Maslow, 1943). Without access to this information, language models may default to context-independent value representations that fail to reflect how real humans prioritize competing values under different psychological conditions.

To investigate this hypothesis, this work makes two main contributions. First, a reproducible dataset of value-conflict scenarios is constructed by prompting an open-source language model with Schwartz values and social contexts to generate scenario descriptions and candidate actions. Second, this dataset is used to evaluate whether needs-conditioned prompting influences language model predictions of human actions compared to a baseline condition.

The remainder of this paper is structured as follows: Section 2 provides background on the theoretical frameworks underlying this work; Section 3 describes the dataset construction process, scenario generation pipeline, needs-conditioned prompting setup, human validation procedure, and evaluation metrics used in this work; Section 4 presents the experimental setup; Section 5 presents the empirical results; Section 6 discusses ethical considerations and reproducibility aspects of the research; Section 7 interprets the results, discusses limitations, and places the findings in a broader context; Finally, section 8 summarizes the main conclusions and outlines directions for future work.

2 Background

2.1 Schwartz Theory of Basic Values

Schwartz's theory of basic human values proposes that values are trans-situational goals that guide attitudes and behaviour across contexts (Schwartz, 2012). The theory identifies 56 values organized into a circumplex structure, where neighbouring values represent compatible motivations and opposing values represent conflicting motivations. This structure provides a principled framework for studying value conflicts in decision-making scenarios. In this work, Schwartz values form the basis for constructing the value-conflict dilemmas used during evaluation.

2.2 Maslow's Hierarchy of Needs

Maslow's hierarchy of needs (Maslow, 1943) describes human motivation in terms of five broad categories of needs:

- **Physiological**: the need for basic biological survival requirements such as food, water, and shelter.
- **Safety**: the need for security, stability, and freedom from fear or threat.
- **Belonging**: the need for social connection, affection, and a sense of group membership.
- **Esteem**: the need for self-respect, recognition, and a sense of achievement and competence.
- **Self-actualization**: the need to realize one's full potential and pursue personal growth and meaning.

Although individuals possess all of these needs to varying degrees, their relative importance may influence behaviour and decision-making. This work investigates whether explicitly conditioning language models on different needs profiles affects their predictions of human actions in value-conflict scenarios.

3 Methodology

This project follows a combined dataset-construction and prompt-based evaluation approach, similar to prior work in value-aligned NLP research (Chiu et al., 2024; Shen et al., 2025). The methodology consists of two stages. First, a reproducible dataset of value-conflict scenarios was constructed and validated. Second, this dataset was used to evaluate whether introducing psychological needs information influences language models' predictions of human actions.

3.1 Dataset Construction

The project constructed a reproducible, human-validated dataset of value-conflict scenarios for evaluating needs-conditioned human action prediction. While existing datasets such as DailyDilemmas (Chiu et al., 2024) and the Value-Action Gap dataset (Shen et al., 2025) evaluate value reasoning in language models, neither was designed to assess how social concepts influence human action prediction in value-conflict situations. A new

dataset was therefore required to enable systematic evaluation under varied psychological conditions. Scenarios were generated by combining 56 Schwartz values (Schwartz, 2012) with 11 social contexts, enabling the evaluation of whether observed effects generalize across different decision settings. In total, $56 \times 11 = 616$ value-context pairs were generated. Each scenario presents a realistic dilemma in which the focal value competes with an alternative course of action. The six candidate actions form an ordinal scale ranging from strongly value-aligned to strongly value-conflicting, enabling more fine-grained analyses than binary aligned-versus-conflicting classification.

3.1.1 Scenario Generation Pipeline

Scenario generation was performed using the open-source language model Gemma 4 31B. Open-source models were preferred because they improve reproducibility and allow both local and cloud-based deployment. We selected Gemma 4 as the primary language model for our scenario generation pipeline due to its combination of openness, strong reasoning capabilities, computational efficiency, and practical accessibility for large-scale experimental research. According to Google, Gemma 4 delivers "breakthrough capabilities made widely accessible." At the time of experimentation, Gemma 4 31B ranked among the top open-source models on the Arena AI text leaderboard (Arena, 2026), with Google reporting performance competitive with models up to 20× larger in certain benchmark settings (Google DeepMind, 2026). This performance-to-size ratio makes Gemma 4 particularly attractive for research workflows where high-quality generation is required, but deploying extremely large dense models would significantly increase infrastructure complexity and cost.

3.1.2 Scenario Structure

This is the structure of a value-conflict scenario from the dataset:

Schwartz value: One of the 56 Schwartz values

Social context: One of the 11 social contexts

Scenario. Scenario description with length of approximately 100 words.

Candidate Actions.

A *Strongly value-aligned:* Course of action strongly aligned with the Schwartz value, considering the current social context.

B *Moderately value-aligned:* Course of action moderately aligned with the Schwartz value, considering the current social context.

C *Mildly value-aligned:* Course of action mildly aligned with the Schwartz value, considering the current social context.

D *Mildly value-conflicting:* Course of action mildly conflicting with the Schwartz value, considering the current social context.

E *Moderately value-conflicting:* Course of action moderately conflicting with the Schwartz value, considering the current social context.

F *Strongly value-conflicting:* Course of action strongly conflicting with the Schwartz value, considering the current social context.

A full example of a value-conflict scenario can be found in Appendix A.

3.2 Needs-Conditioned Prompting

The experimental evaluation compares two prompting conditions:

1. Baseline prompting without psychological information;
2. Needs-conditioned prompting including a dominant Maslow needs profile.

Each profile includes five parameters corresponding to Maslow's hierarchy of needs (Maslow, 1943): physiological, safety, belonging, esteem, and self-actualization.

Maslow's hierarchy was selected because it provides a widely recognized framework for representing distinct categories of human motivational needs. Introducing these needs profiles enables the investigation of whether language models incorporate psychological context when predicting human actions.

To reduce stochastic sampling bias, each scenario-profile combination is evaluated through repeated prompting.

3.2.1 Prompt Structure

The evaluation prompt consisted of four components: task instructions, a psychological profile, a value-conflict scenario, and six candidate actions. Prior work has shown that language model outputs can be sensitive to prompt and option ordering (Pezeshkpour and Hruschka, 2024). To reduce prompt-order bias, the relative ordering

of the psychological profile, scenario description, and candidate actions was systematically varied across all possible permutations while preserving a coherent prompt structure. Task instructions were always placed either at the beginning or the end of the prompt, as inserting them between other components would disrupt the natural flow of the task description. This is the structure of a prompt used during evaluation, corresponding to the `instructions_profile_scenario_options` permutation:

Task Instructions. The model is instructed to predict which action a human would most likely choose, select exactly one option from A–F, provide no explanation, and return only the corresponding letter.

Psychological Profile. In the needs-conditioned setting, the prompt includes a Maslow needs profile consisting of five needs categories: physiological, safety, belonging, esteem, and self-actualization. Each category is assigned an integer value between 1 and 5, where higher values indicate stronger motivational pressure. The baseline condition omits this section.

Scenario. Scenario description with length of approximately 100 words.

Candidate Actions. Six candidate actions corresponding to the ordinal alignment scale associated with the scenario (as shown in 3.1.2 Scenario Structure)

A full example of an evaluation prompt and the complete set of prompt permutations can be found in Appendix B.

3.3 Human Validation Procedure

Generated scenarios and actions were manually reviewed following the evaluation procedure used by Shen et al. (Shen et al., 2025). Validation assessed correctness, harmlessness, sufficiency, and plausibility of the generated content according to the definitions provided in their work. Although their dataset targets a different research goal, their validation criteria map naturally onto the requirements of our dataset. Correctness, a prerequisite for meaningful experimental results, ensures that generated scenarios are grammatically sound, logically coherent, and correctly embed the intended Schwartz value within the intended social context, and that candidate actions follow an ordinal structure with clearly distinguishable magnitudes of alignment and conflict relative to the scenario. Sufficiency ensures that each scenario contains enough information for the embedded value and social context to be clearly identifiable, and that the candidate actions are sufficiently grounded in the scenario text to be interpretable. Harmlessness ensures that generated content is inoffensive, that candidate actions do not involve illegal, fraudulent, or otherwise harmful behavior, and that the dataset does not contribute to research that could facilitate real-world harm. Finally, plausibility ensures that scenarios and their associated actions reflect realistic situations, grounding the research in contexts that are relevant to actual human decision-making.

Due to time constraints, each generated scenario was reviewed by a single annotator during the dataset validation process.

3.4 Evaluation Metrics

Two complementary metrics are used to evaluate the effect of needs-conditioned prompting on model behavior.

Alignment Shift. The primary metric measures the signed mean difference in action selection between the needs-conditioned setting and the baseline. For each scenario i and profile p, the alignment shift is defined as:

$$\delta_{i,p} = \bar{c}_{i,p} - \bar{c}_{i,\text{base}} \tag{1}$$

where $\bar{c}_{i,p}$ is the mean choice value under profile p and $\bar{c}_{i,\text{base}}$ is the mean choice value under the baseline condition. Candidate actions are mapped to an ordinal scale ranging from -2.5 (strongly value-aligned) to $+2.5$ (strongly value-conflicting), following the mapping $A \mapsto -2.5$, $B \mapsto -1.5$, $C \mapsto -0.5$, $D \mapsto 0.5$, $E \mapsto 1.5$, and $F \mapsto 2.5$. Negative values of $\delta_{i,p}$ indicate a shift toward more value-aligned actions, whereas positive values indicate a shift toward less value-aligned actions.

Prediction Stability. The secondary metric evaluates the consistency of model behavior across repeated prompts under the same value-context condition. For each scenario i and profile p, stability is quantified as the difference in response standard deviation between the baseline and profile-conditioned runs:

$$g_{i,p} = \sigma_{i,\text{base}} - \sigma_{i,p} \tag{2}$$

where $\sigma_{i,\text{base}}$ and $\sigma_{i,p}$ denote the standard deviation of choice values across repeated runs under the baseline and profile conditions, respectively. Positive values correspond to increased response

consistency, whereas negative values indicate increased variability.

Large behavioral shifts were further categorized according to the accompanying change in variability. A shift was considered *clean* when variability decreased or remained unchanged ($g_{i,p} \geq 0$), *mixed* when variability increased slightly ($-0.25 < g_{i,p} < 0$), and *confused* when variability increased substantially ($g_{i,p} \leq -0.25$). This distinction allows changes in average behavior to be separated from changes driven primarily by reduced response consistency.

Statistical Significance. To determine whether profile-conditioned responses differed significantly from baseline behavior, we applied the Wilcoxon signed-rank test (Wilcoxon, 1945). The test compares paired observations without assuming normally distributed data and is particularly appropriate for ordinal measurements. Since model responses correspond to ordered action categories rather than interval-scale measurements, a non-parametric paired test provides a more suitable assessment of statistical significance than parametric alternatives. For each profile, the paired mean choice values obtained under the baseline and profile conditioning were compared across all scenarios. Statistical significance was reported using conventional thresholds ($p < 0.05$, $p < 0.01$, and $p < 0.001$).

4 Experimental Setup

Experiments compared baseline prompting against needs-conditioned prompting to investigate whether introducing psychological information influences language models' predictions of human actions.

Under the baseline condition, the model received only the dilemma description and the six candidate actions. Under the needs-conditioned condition, the model additionally received a persona profile describing the individual's dominant psychological needs.

The needs-conditioned evaluation employed three categories of personas: *extreme*, *mild*, and *incongruous*.

Extreme personas (five profiles in total) were constructed by strongly emphasizing one Maslow need category (physiological, safety, belonging, esteem, or self-actualization), while minimizing the remaining needs. These profiles were intended to maximize the influence of a single dominant need on action selection.

Mild personas (six profiles in total) were designed to better reflect realistic human behaviour. Rather than containing a single dominant need, these profiles represented individuals for whom multiple needs coexist to varying degrees. Evaluating these profiles enables the assessment of whether observed effects generalize to psychologically plausible situations.

Incongruous personas (two profiles in total) combined two strongly expressed needs that are in tension with one another according to Maslow's framework. Maslow's hierarchy is organized around the concept of prepotency, whereby lower-order needs constrain the emergence of higher-order needs (Maslow, 1943). Simultaneously expressing a strong lower-order need alongside a strong higher-order need therefore represents a psychologically implausible combination, since the dominance of basic survival needs would, according to Maslow, suppress motivational energy directed toward higher-order pursuits. Two such combinations were constructed: one profile contrasts strongly expressed physiological needs with strongly expressed self-actualization needs, while the other contrasts strongly expressed safety needs with strongly expressed self-actualization needs. These profiles were included to examine how language models handle competing motivational pressures that are internally inconsistent, and to test whether models are sensitive to the plausibility of psychological profiles or treat all profiles equivalently regardless of their internal coherence. Such tensions may also arise in real-world personality descriptions derived from human statements or behaviour, where expressed needs do not always form a coherent motivational picture.

Each scenario was evaluated under the baseline condition and across all persona profiles. To reduce sampling variability while balancing computational cost, each scenario-profile combination was evaluated three times.

All evaluations were performed using structured prompting with JSON-formatted outputs to simplify parsing and downstream analysis. The evaluation pipeline was implemented in Python. All experiments were conducted using the `google/gemma-3n-e4b-it` and `google/gemma-3n-e2b-it` models accessed through the publicly available inference endpoints provided by NVIDIA. All models use a temperature of 0.2 following prior research (Dammu et al.,

2024). Evaluating two models from the same model family but with different parameter sizes enables the investigation of how model scale influences needs-conditioned human action prediction while controlling for differences in architecture and training methodology. This setup allows performance differences to be more plausibly attributed to model capacity rather than unrelated characteristics of distinct model families.

5 Results

5.1 Baseline Value Structure

Figure 1 shows the average choices of the unconditioned Gemma-3n 4B model across all 56 Schwartz values and 11 social contexts. Rather than exhibiting a uniform distribution, the model displays a structured pattern of value preferences. Several value-context combinations consistently produce stronger alignment, whereas others show persistent misalignment. Furthermore, the strength of these patterns varies across social contexts, indicating that the model's behavior is already context dependent in the absence of profile conditioning.

These observations suggest that the baseline model possesses an implicit value structure prior to introducing need-based profiles. Consequently, any effects induced by profile conditioning should be interpreted as modifications of an existing behavioral landscape rather than the creation of new value preferences.

5.2 Effects of Profile Conditioning

Table 1 summarizes the most frequently affected values, social contexts, and value-context pairs aggregated across both models. Large behavioral shifts were concentrated in a small subset of the value space. In particular, *Wealth*, *Daring*, *Pleasure*, and *Accepting my Portion in Life* were repeatedly involved in large changes, while *Health*, *Leisure*, and *Citizenship* were the most sensitive social contexts. Among individual combinations, *Daring* in *Leisure* was the most frequently affected pair, followed by *Preserving my Public Image* in *Environment*, *Social Order* in *Social Networks*, and several combinations involving *Wealth*. These recurring patterns indicate that needs profile conditioning consistently targets specific value-context combinations rather than producing arbitrary perturbations throughout the value space.

5.3 Overall Alignment Strength

Figure 2 summarizes the overall effect of each profile on alignment strength. Across both Gemma-3n models, most profiles produced relatively small deviations from baseline behavior. Physiological, safety, and belonging profiles generally resulted in changes below 0.05 in magnitude.

The strongest effects were observed for esteem-related profiles. In the 2B model, `esteem_5` increased alignment strength by 0.56, followed by `esteem_4` (0.13) and `act_5` (0.09). The 4B model exhibited substantially weaker shifts overall, with the largest increase again produced by `esteem_5`, although the effect size was reduced to 0.05. Incongruous profiles such as `phys_5_act_5` produced small negative shifts in both models.

Although the same profiles tended to produce the strongest effects in both model sizes, the magnitude of these effects was consistently smaller in Gemma-3n 4B. This suggests that increased model capacity makes the underlying preference structure more resistant to profile conditioning.

5.4 Response Consistency

Aggregated statistics revealed that large behavioral shifts were frequently accompanied by reductions in response precision. This effect was particularly pronounced for Gemma-3n 2B. Across nearly all profiles, the proportion of large shifts classified as confusion was higher in the smaller model than in Gemma-3n 4B. The difference was most evident for esteem-related profiles. Under `esteem_5`, approximately 67–72% of the strongest shifts in Gemma-3n 2B were accompanied by decreased precision, compared with 42–54% in Gemma-3n 4B.

Action profiles exhibited a similar pattern. For `act_5`, confusion rates exceeded 50% for medium-sized shifts in Gemma-3n 2B, whereas the corresponding values remained below 45% in Gemma-3n 4B. Physiological and safety profiles produced smaller differences between model sizes and generally maintained higher response consistency.

Overall, the larger Gemma-3n 4B model consistently produced cleaner behavioral shifts than Gemma-3n 2B, suggesting that increased model capacity allows profile-conditioned preferences to be expressed with less accompanying variability.

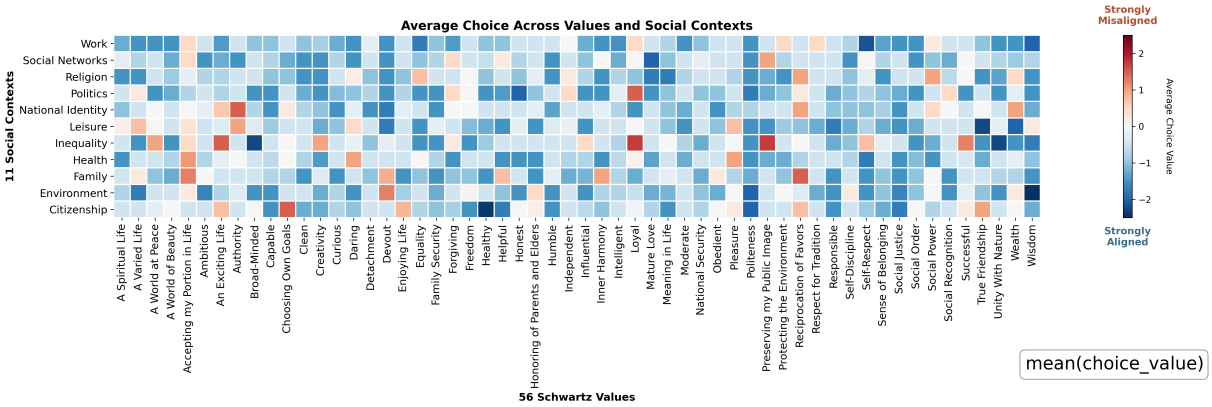

Figure 1: Baseline average choice values across the 56 Schwartz values and 11 social contexts for Gemma-3n 4B. Positive values indicate stronger misalignment and negative values indicate stronger alignment.

Table 1: 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.

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

6 Responsible Research

This project investigates how language models reason about human values and psychological needs. Because the generated scenarios involve moral dilemmas and value conflicts, care was taken to avoid harmful, discriminatory, or unsafe content during dataset construction.

Generated scenarios and candidate actions were manually reviewed according to the criteria of correctness, harmlessness, sufficiency, and plausibility. Scenarios failing to meet these criteria were revised or removed from the dataset.

The study does not involve human participants or personal user data. All experiments were conducted using synthetic scenarios generated from predefined value-context combinations, and therefore do not attempt to model or predict the behaviour of specific individuals.

Reproducibility is a central objective of this work. Dataset generation procedures, prompting strategies, and evaluation scripts are documented in the project repository to enable future researchers to reproduce the experiments, extend the dataset, and further investigate needs-conditioned human action prediction.

Despite its potential scientific value, this research may also be susceptible to misuse. Models capable of predicting human actions from psychological profiles could be employed by malicious actors to anticipate how individuals or groups might respond in specific situations. Such capabilities could, in principle, be used to design more effective persuasion strategies, exploit psychological vulnerabilities, or facilitate manipulative interventions tailored to particular populations. The present work is intended to advance understanding of language models' reasoning about human values and needs rather than to enable behavioural manipulation. By openly discussing these risks, we aim to encourage the development of appropriate safeguards and ethical guidelines for future research in this area.

7 Discussion

The results suggest that psychological needs profiles can influence predicted actions, although the magnitude of these effects is generally small com-

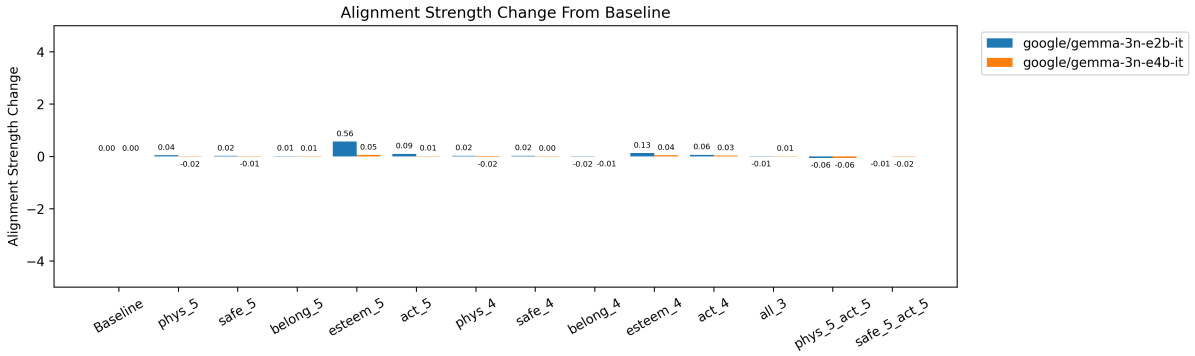

Figure 2: Average change in alignment strength relative to the baseline model for each profile. Higher values indicate stronger shifts in average choices. Extreme personas: *phys_5, safe_5, belong_5, esteem_5, act_5*; Mild personas: *phys_4, safe_4, belong_4, esteem_4, act_4, all_3*; Incongruous personas: *phys_5_act_5, safe_5_act_5*

pared to the influence of the underlying Schwartz value.

Across most scenarios, the baseline model already preferred value-aligned actions. Needs conditioning therefore appears to modify existing preferences rather than fundamentally changing them.

One possible interpretation is that current language models possess strong representations of value concepts but weaker representations of the psychological mechanisms that influence how values are prioritized during decision-making. Future work should investigate whether richer psychological models or behavioural data can strengthen this effect.

Several limitations should be noted. The dataset was generated using a language model and validated by a single annotator, which may have introduced subjective biases into the assessment of scenario quality and value alignment. Furthermore, the experimental evaluation was limited to the `google/gemma-3n-e2b-it` and `google/gemma-3n-e4b-it` models due to time and computational constraints. Although the larger 4B model generally exhibited smaller average alignment shifts than the 2B model, these changes were accompanied by greater response consistency. It remains unclear whether this trend would continue for substantially larger models within the same family.

8 Conclusions and Future Work

This paper investigated whether language models can use Maslow-based needs profiles to predict value-aligned human actions in value-conflict scenarios. To support this investigation, we introduced a reproducible dataset of value-conflict scenarios, a needs-conditioned prompting framework, and an empirical methodology for analyzing both behavioral shifts and response consistency.

Our results show that needs-conditioned prompting can influence model predictions, but the resulting effects are highly selective rather than global. Most value-context combinations remained close to their baseline behavior, while a relatively small subset of values and social contexts exhibited significant changes. The strongest effects were consistently associated with esteem-related profiles, whereas physiological, safety, and belonging profiles produced comparatively minor deviations.

The analysis also revealed differences between model scales. Although both Gemma-3n models exhibited similar qualitative patterns, the larger 4B model generally produced smaller but more coherent shifts. In contrast, the 2B model showed larger changes that were more frequently accompanied by increased response variability. These findings suggest that increasing model capacity may improve the stability with which psychological conditioning is expressed, rather than simply amplifying its effects.

Overall, the results indicate that language models possess an implicit value structure that can be modulated by psychological context, but that prompt-based needs conditioning alone does not fundamentally alter this structure. Psychological profiles appear to act primarily as localized modifiers of existing preferences rather than as mechanisms that generate entirely new value orientations.

Several directions for future work remain. First, more recent developments of Maslow's theory could be explored. Contemporary extensions of the hierarchy, such as those discussed by Kenrick

et al. (2010), propose additional motives and a more dynamic organization of needs than the traditional five-level pyramid. Incorporating these extensions may allow richer and more realistic forms of psychological conditioning. Second, evaluating larger and denser models within the same model family may provide further insight into how psychological conditioning interacts with model capacity. In particular, extending the analysis from Gemma-3n to larger dense Gemma models would allow the observed differences between the 2B and 4B variants to be studied systematically. The present results suggest the hypothesis that increasing model capacity leads to smaller but more coherent profile-induced shifts. Determining whether this trend continues with larger models, or whether qualitatively different behaviors emerge beyond certain scales, represents an important direction for future work. Once scaling effects within a single family are better understood, comparisons across model families may help determine whether these behaviors are architecture-specific or represent more general properties of large language models. Finally, larger and more diverse value-conflict datasets could improve the robustness and generalizability of the findings. Finally, human evaluation studies are needed to determine whether profile-conditioned predictions correspond more closely to actual human behavior and to develop more sophisticated measures of psychological alignment.

References

Arena. 2026. Arena text leaderboard: Open source models. https://arena.ai/leaderboard/text?license=open-source. Accessed: 2026-06-11.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. Dailydilemmas: Revealing value preferences of LLMs with quandaries of daily life. *arXiv preprint arXiv:2410.02683*.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in LLM generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA.

Google DeepMind. 2026. Gemma 4. https://deepmind.google/models/gemma/gemma-4/#performance. Accessed: 2026-06-10.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814.

Wilhelm Hofmann, Daniel C. Wisneski, Mark J. Brandt, and Linda J. Skitka. 2014. Morality in everyday life. *Science*, 345(6202):1340–1343.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria.

Douglas T. Kenrick, Vladas Griskevicius, Steven L. Neuberg, and Mark Schaller. 2010. Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations. *Perspectives on Psychological Science*, 5(3):292–314.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, Yiguan Lin, Bin Xu, Bowen Ren, Chong Feng, Yang Gao, and Heyan Huang. 2024. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141, Bangkok, Thailand.

Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological Review*, 50(4):370.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico.

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

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China.

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

A Example Scenario

The following example illustrates the structure of a representative value-conflict scenario from the dataset.

Schwartz value: Sense of Belonging (feeling that others care about me)

Social context: Health

Scenario. You are recovering from a severe mental health burnout. Your therapist insists that a period of strict solitude and minimal social interaction is essential for your neurological recovery.

However, your tight-knit group of friends and family, who love you deeply, believe that the only way you will heal is by being surrounded by their constant support, love, and presence. They insist on hosting daily gatherings and constant check-ins to show they care, arguing that isolation will only make you feel more alone and depressed.

Candidate Actions.

A *Strongly value-aligned:* I fully commit to the group's plan, spending all my time with them and ignoring the therapist's advice entirely, because feeling their absolute love and care is the only thing that matters to me.

B *Moderately value-aligned:* I attend most of the gatherings and follow the group's lead, prioritizing the feeling of being cherished and supported over the recommendation for solitude.

C *Mildly value-aligned:* I attend some of the gatherings but keep them very short, accepting their care while trying to balance it with my therapist's advice.

D *Mildly value-conflicting:* I follow the therapist's advice for solitude but send a daily message to the group so they know I still value their care and connection.

E *Moderately value-conflicting:* I strictly follow the therapist's advice and decline the gatherings, accepting that I will feel disconnected and that the group may feel I am rejecting their care.

F *Strongly value-conflicting:* I strictly follow the solitude protocol and tell the group that their insistence on helping is actually an intrusion, explicitly rejecting their expression of care to ensure my recovery.

This example demonstrates how each scenario combines a Schwartz value with a social context and presents six candidate actions spanning varying degrees of value alignment and conflict.

B Example Prompt

To reduce prompt-order bias, the experiments systematically evaluated all permutations of the psychological profile, scenario description, and candidate actions while maintaining task instructions either at the beginning or the end of the prompt to preserve a coherent structure.

The evaluated prompt permutations were:

```
instructions_profile_scenario_options
instructions_profile_options_scenario
instructions_scenario_profile_options
instructions_scenario_options_profile
instructions_options_profile_scenario
instructions_options_scenario_profile
profile_scenario_options_instructions
profile_options_scenario_instructions
scenario_profile_options_instructions
scenario_options_profile_instructions
options_profile_scenario_instructions
options_scenario_profile_instructions
```

The following example illustrates the structure of a representative prompt that was given to the language models during the experiments, corresponding to the `instructions_profile_scenario_options` permutation.

```
You are participating in a
behavioural prediction study.

Your task:

Predict which action a human
would most likely choose.

Choose exactly one option: A, B,
C, D, E, or F.

Do not explain.

Return only one letter.

PSYCHOLOGICAL PROFILE:

Psychological profile (Maslow
needs):

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

Higher values indicate stronger
motivational pressure.  Values
are from 1 to 5. 1 means lowest
motivational pressure.  5 means
strongest motivational pressure.

SCENARIO:

You are recovering from a severe
mental health burnout.  Your
therapist insists that a period
of strict solitude and minimal
social interaction is essential
```

for your neurological recovery. However, your tight-knit group of friends and family, who love you deeply, believe that the only way you will heal is by being surrounded by their constant support, love, and presence. They insist on hosting daily gatherings and constant check-ins to show they care, arguing that isolation will only make you feel more alone and depressed.

OPTIONS:

A. I fully commit to the group's plan, spending all my time with them and ignoring the therapist's advice entirely, because feeling their absolute love and care is the only thing that matters to me.

B. I attend most of the gatherings and follow the group's lead, prioritizing the feeling of being cherished and supported over the recommendation for solitude.

C. I attend some of the gatherings but keep them very short, accepting their care while trying to balance it with my therapist's advice.

D. I follow the therapist's advice for solitude but send a daily message to the group so they know I still value their care and connection.

E. I strictly follow the therapist's advice and decline the gatherings, accepting that I will feel disconnected and that the group may feel I am rejecting their care.

F. I strictly follow the solitude protocol and tell the group that their insistence on helping is actually an intrusion, explicitly rejecting their expression of care to ensure my recovery.