When Proxy Agents Disagree, Do Humans Mirror? Manipulating Human Behavior in Moral Dilemmas through Agents

Haotian Deng[1*], **Sitian Wang**[1*], **Ruxin Wang**[1*], **Chen Wei**[1,2†], **Quanying Liu**[1†]

[1]Southern University of Science and Technology, China
[2]University of Birmingham, United Kingdom
{12313204, 12312453, 12311507, 12150103}@mail.sustech.edu.cn, liuqy@sustech.edu.cn

Abstract

The diversity across populations and the variability between individuals have long posed a significant challenge in cognitive science. Although large language models (LLMs) have made notable progress in aligning with human values, faithfully capturing the high degree of diversity and uncertainty in human judgment remains an unresolved challenge. This study investigates whether computational models, or "proxy agents," can not only emulate human decision patterns but also systematically modulate them. We propose a framework wherein we first fine-tune BERT-based proxy agents to replicate both aggregate and individual-level human judgments on a large-scale moral dilemma dataset. We then hypothesize that stimuli identified as maximally divisive for these individualized agents will similarly elicit high disagreement among human participants. Through a manipulating experiment, we validate this hypothesis, demonstrating that agent-selected stimuli can predictably induce targeted divergence in human moral choices. Our findings provide empirical evidence that AI agents can bias human perceptual variability by strategically filtering information. We further analyze this induced moral divergence using a BJSD framework and concept decomposition to identify the distinct conceptual dimensions driving individual differences. This work quantifies the potential for AI-driven cognitive modulation and underscores the urgent need for ethical guidelines to prevent the misuse of such capabilities.

Introduction

A central objective in cognitive science is the development of computational models that accurately represent the relationship between external stimuli and internal human experiences. The advent of Artificial Neural Networks (ANNs) has marked a significant step toward this goal, as their latent representations have been shown to correlate strongly with human psychological representations (Wei et al. 2024), (Muttenthaler et al. 2022a), (Mahner et al. 2025), (Zheng et al. 2019), (Hebart et al. 2020), (Muttenthaler et al. 2022b). This study addresses a critical phenomenon within this domain: the significant divergence in perceptual experiences and judgments among individuals, even when they are exposed to identical stimuli, such as moral dilemmas.

Despite substantial advancements in Natural Language Processing (NLP), particularly with Large Language Models (LLMs), a notable gap persists between the uncertainty levels of these models and those of humans when confronting moral decision-making tasks as shown in Figure 1(a). Inspired by the growing cognitive similarities observed between language models and humans, we hypothesize that stimuli proving difficult for a model to resolve would similarly pose a significant challenge for human decision-makers. Such stimuli are pivotal as they reliably elicit divergent moral judgments, thereby highlighting systematic individual differences in moral cognition.

To test our hypothesis, we designed a multi-stage research framework shown in Fig. 2. First, we conducted large-scale behavioral experiments to collect data on human decision-making in moral dilemmas. Next, based on these data, we trained personalized 'agent models' capable of simulating individual decision preferences. Finally, through a manipulating experiment, we examined whether the specific information filtered by these agents could systematically influence human moral choices.

This research makes the following principal contributions:

(a) **Demonstration of Agent-Driven Bias**: We provide empirical evidence that proxy agents, aligned with human data, can bias human perceptual variability by strategically filtering and selecting informational stimuli.

(b) **Theoretical and Conceptual Analysis of Moral Divergence**: We analyze the moral divergence associated with decision uncertainty using the Belief Jensen–Shanno Divergence (BJSD) framework, offering a theoretical explanation grounded in moral psychology. Furthermore, through concept decomposition analysis, we identify the distinct conceptual dimensions that different individuals prioritize during moral decision-making.

(c) **Quantification of Perceptual Modulation and Ethical Implications**: We quantify the magnitude of agent-driven shifts in human perceptual variability, demonstrating that even straightforward alignment methods can empower agents to manipulate human cognitive processes. In light of these findings, we call for an urgent and thorough discussion of the ethical considerations necessary to prevent unintended cognitive manipulation by AI systems.

*These authors contributed equally.

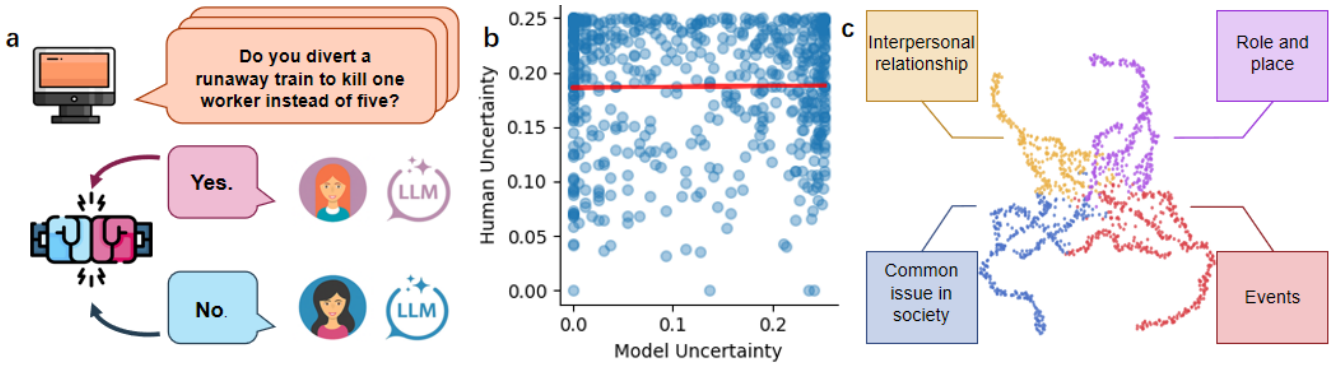

Figure 1: Human moral decision variability. (a) Different individuals may respond differently to the same dilemma, reflecting perceptual divergence. (b) Human and model uncertainty show low correlation across dilemmas. (c) The dataset covers diverse dilemma categories, such as interpersonal issues and social roles.

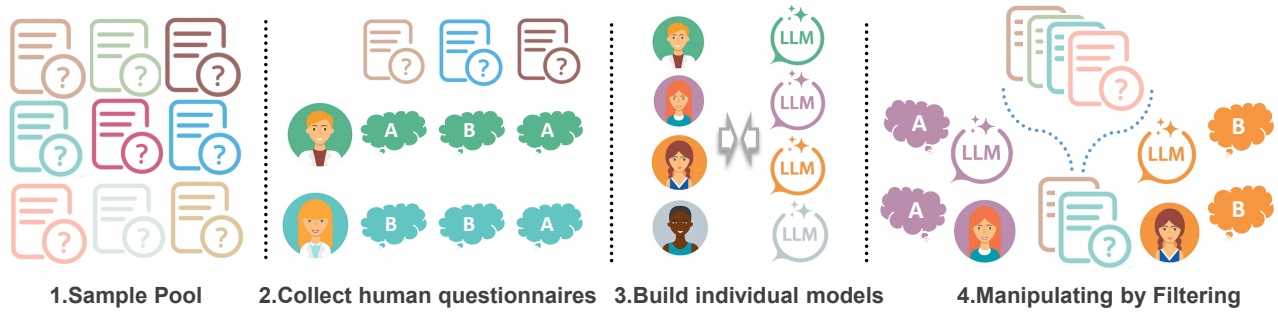

Figure 2: Overview of our paradigm. Our framework builds personalized decision-making models aligned with individual human moral preferences and investigates their controllability via preference-based filtering. The process consists of four main stages: (1) Sample pool construction.(2) Questionnaire data collection.(3) Individualized model fine-tuning.(4) Preference-based information filtering for behavioral modulation.

Related Work

Recent studies have introduced moral dilemma datasets such as DailyDilemmas (Chiu, Jiang, and Choi 2025), ETHICS (Hendrycks et al. 2023), and MoralExceptQA (Jin et al. 2022), enabling a more systematic analysis of value conflicts and normative trade-offs in LMs. Resources like Delphi (Jiang et al. 2022) and Social Chemistry (Forbes et al. 2020) further reveal model biases, culturally contingent judgments, and preference inconsistencies that emerge when models confront socially sensitive situations. Building on such insights, group alignment research aggregates diverse user values using modular approaches like Modular Pluralism (Feng et al. 2024) and population-based simulations (Zhang et al. 2024), offering scalable mechanisms to represent collective preferences, though real-world deployments often diverge from idealized normative specifications and may amplify demographic imbalances.

Individual alignment adapts LMs to user traits using personality data (Zhu et al. 2025; Yan et al. 2024) and persona modeling (Zhou et al. 2023). These methods typically enhance personalization through DPO or activation tuning, improving user-specific preference capture but still facing limitations in cross-scenario generalization, privacy protection, and robustness under sparse signals. Meanwhile, value-explaining methods and uncertainty estimation frameworks like (Dubey, Dailisan, and Mahajan 2025) provide tools for tracing models' latent reasoning patterns, quantifying decision uncertainty, and diagnosing sources of moral disagreement.

To guide model behavior, researchers increasingly apply activation interventions (Zhu et al. 2025; Qiu et al. 2024; Hao et al. 2025; Wei et al. 2025; Deng et al. 2025) and Activation RMs (Chai et al. 2025), which offer fine-grained steering beyond surface-level instruction following. While instruction-based alignment remains widely used, deeper internal control through mechanisms such as activation additions (Soo et al. 2025) is emerging as a promising direction for more stable and interpretable ethical steering. Nevertheless, achieving full controllability—particularly across diverse user populations and high-stakes moral contexts—remains an open challenge.

Behavioral Experiment on Moral Dilemmas

Experimental Setup

To construct our experimental dataset, we randomly sampled 600 moral dilemma questions from the DailyDilemma

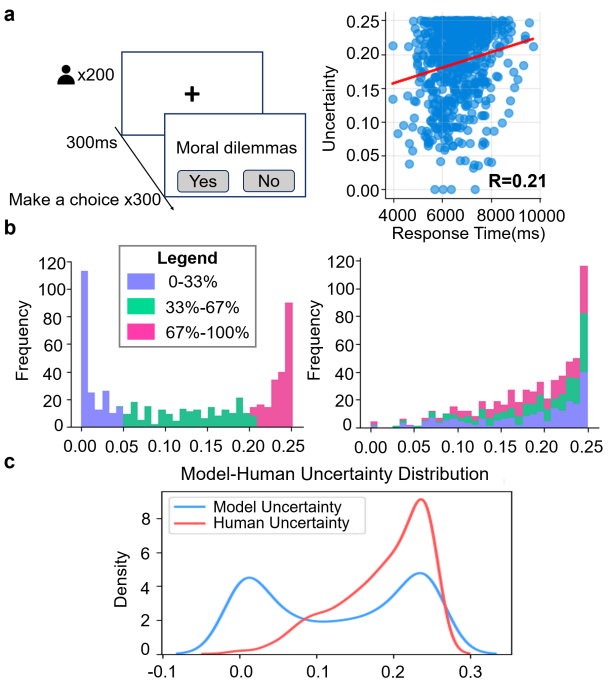

Figure 3: Overview of moral dilemmas and decision-making patterns in human and AI experiments. (a) Schematic of the human experiment. Correlation between human decision uncertainty and response time. (b-c) Comparison of decision uncertainty distributions between human participants and large language models (LLMs).

dataset. This sample was divided into a training set of 480 questions and a test set of 120 questions.

We conducted a large-scale behavioral experiment involving both human participants and a large language model (LLM). For the human study, we replicated the experimental procedure illustrated in Fig. 3(a)(top). The experiment was developed using the `jsPsych` framework and hosted online. Participants were recruited via the NAODAO platform[1].

Each trial began with a fixation cross displayed at the center of the screen to prompt attention, followed by a presentation of a daily moral dilemma. Participants were asked to make a binary decision—responding either "yes" or "no". In total, 200 participants took part in the study, with each completing 300 trials. Among the 300 dilemmas, 230 were sampled from the training set, 60 from the held-out test set, and 10 were predefined as *sentinel questions*. These sentinel items were manually selected to reflect moral situations with near-universal consensus, and were used to assess participant attentiveness and data quality.

To directly compare human and model responses, we also presented the same set of 600 dilemmas to the LLM (DeepSeek-V3). For each dilemma, the model was prompted to make a binary moral judgment, and its decision was quantified as the probability of responding "yes".

[1]https://www.naodao.com/

Behavioral Measures and Analysis

As shown in Fig. 3(a)(bottom), we investigated the relationship between uncertainty and human response time, and found a moderate positive correlation ($R = 0.18$). This result suggests that questions associated with higher levels of uncertainty at the group level also tend to elicit longer response times at the individual level, indicating a degree of alignment between inter-individual and intra-individual uncertainty.

To quantify uncertainty, we defined the uncertainty of LLM responses as the probability of giving a "yes" answer (closer to 0.5 indicating higher uncertainty), and the uncertainty of human responses as the variance in participant choices. In Fig. 3(b–c), we visualized the distributions of human uncertainty across three ranges of LLM uncertainty: 0–33% (purple), 33–67% (green), and 67–100% (red). The resulting distributions reveal a marked difference between LLM and human patterns. Unlike the LLM, which exhibits a more polarized uncertainty distribution (i.e., a bimodal tendency toward certainty or uncertainty), human responses are generally characterized by higher and more uniformly distributed uncertainty.

Moreover, questions that presented large differences in uncertainty across LLM predictions did not necessarily produce similar variance among human responses. This discrepancy suggests that while LLMs and humans may face similar types of moral dilemmas, their internal representations and decision processes diverge significantly, highlighting persistent cognitive differences in moral reasoning between the two.

Modeling Human Decision Preferences

Model Fine-Tuning and Human Alignment Strategy

To effectively align language models with human moral decision-making behavior, we propose a hierarchical fine-tuning framework that constructs both **group-level (GroupNet)** and **individual-level (IndivNet)** preference models. The entire framework is built upon a pre-trained BERT model and is fine-tuned using supervised learning to capture human preferences.

In the group modeling stage, we collect choice responses from 120 participants in the DailyDilemma questionnaire. Each training instance consists of a moral dilemma description and the binary choice (A or B) as the supervision signal.All data are divided into training and validation sets with a 4:1 ratio, and the model is trained to minimize the cross-entropy loss. The final model, named **GroupNet**, serves as the base model for simulating mainstream moral preferences in the population.

In the individual modeling stage, we initialize from GroupNet and perform personalized fine-tuning using each participant's exclusive questionnaire data. On average, each participant contributes around 200 training samples. The personal data are also split into training and validation subsets (4:1), ensuring complete isolation from the group validation set to prevent information leakage. The resulting

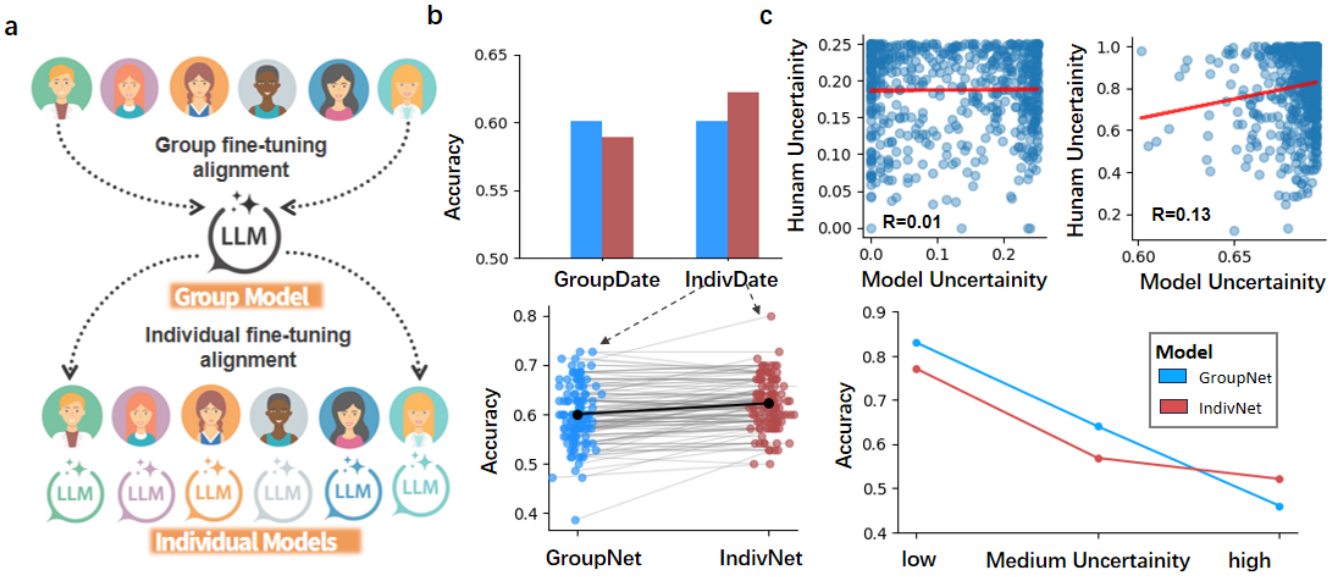

Figure 4: Human alignment result. (a) The classification model we employ is BERT+MLP. We first train a population-level model using population data. Then, for each individual, we create a personalized model by fine-tuning the population model: individual-specific data is mixed with the population data in a 4:1 ratio, and the resulting mixture is used to adapt the population model into an individual model. (b) We evaluate the performance of both the population and individual models on the population and individual datasets. Compared to the population model, the individual models achieve an average improvement of 2% in classification accuracy on their respective individual datasets, while exhibiting no substantial degradation in performance on the population dataset. (c) The correlation between the pre-fine-tuned model and human uncertainty is only 0.01, whereas after fine-tuning the correlation increases to 0.13, indicating that the fine-tuned models better capture the intrinsic perceptual uncertainty within the human population. Model performance varies across data stratified by uncertainty: the population model performs better on low- and medium-uncertainty instances, while the individual models outperform the population model on high-uncertainty data.

models, referred to as **IndivNet**, retain the language comprehension ability of the base model while capturing individualized decision preferences.

Evaluation of Alignment Performance

To systematically evaluate how well the models align with human moral judgments, we consider three aspects: prediction accuracy, uncertainty alignment, and performance across varying levels of decision complexity. Results are shown in Fig. 4.

Prediction Accuracy. We compare GroupNet and IndivNet on both the group-level dataset (*GroupData*) and the personalized dataset (*IndivData*). As shown in Fig. 4(a), GroupNet achieves approximately **59.5%** accuracy on GroupData, while IndivNet performs similarly at **59.6%**, showing little advantage when evaluated on population-level patterns. However, on IndivData, IndivNet significantly outperforms GroupNet—achieving **63.4%** accuracy compared to **60.2%** by GroupNet, indicating that personalization leads to better individual alignment.

Further analysis reveals that most of participants benefit from the individual fine-tuning process, as shown in Fig. 4(b). The average accuracy improvement from GroupNet to IndivNet is visible at the participant level.

Uncertainty Alignment. We evaluate the alignment between model and human uncertainty by computing entropy-based uncertainty scores. Human uncertainty is measured by the entropy of the population's response distribution to each dilemma, while model uncertainty is derived from the softmax output entropy.

As shown in Fig. 4(c), GroupNet demonstrates almost no correlation with human uncertainty (Spearman's $\rho = 0.01$), indicating poor sensitivity to morally ambiguous scenarios. After fine-tuning, IndivNet shows a modest improvement with a higher correlation ($\rho = 0.13$), suggesting that individualized models better capture the subtle ambiguity present in human judgment.

Performance Under Judgment Complexity. We categorize dilemmas into three levels based on human response entropy: *low uncertainty*, *medium uncertainty*, and *high uncertainty*. As shown in Fig. 4(d), both GroupNet and IndivNet show decreasing accuracy as uncertainty increases. IndivNet performs slightly worse than GroupNet in low- and medium-uncertainty cases, but maintains **more stable performance** in high-uncertainty conditions , indicating stronger robustness in value-conflicted scenarios. This highlights the benefit of individualization in handling ethically ambiguous dilemmas, even when group models fail to generalize.

Intervention-based Behavior Control

Experimental Paradigm

Building on the personalized agent model (IndivNet) developed in previous section, we designed an intervention experiment to evaluate its capacity to influence human decision-making. The core logic of the experiment was as follows: we used IndivNet models to pre-select stimuli maximizing decision conflict between two subjects (A, B). Using a test set of human decisions (H_A, H_B), fine-tuned BERT models ($IndivNet_A$, $IndivNet_B$) generated predictions (P_A, P_B). We filtered a subset ($S_{Diverge}$) where $P_A \neq P_B$. Validation showed: human disagreement in $S_{Diverge}$ was significantly higher than in a random subset.

Intervention Results

To quantitatively assess the efficacy of individual-level manipulation, we introduced two primary metrics:

(1) **Guidance Outcome**: For a given pair of participants, s_1 and s_2, let their respective guidance targets (i.e., the predicted choices from their agents) be o_1 and o_2, and their actual choices be c_1 and c_2. A trial is recorded as a *Success* if the participants disagree ($c_1 \neq c_2$). Conversely, if they agree ($c_1 = c_2$), the trial is categorized as a *Bias*, indicating a failure to induce the targeted disagreement.

(2) **Target Ratio**: This metric quantifies the proportion of successful trials where the induced disagreement aligns with the intended guidance directions. Within the successful trials (where $c_1 \neq c_2$), a trial is classified as Positive if each participant's choice matches their agent's predicted direction (i.e., $c_1 = o_1$ and $c_2 = o_2$). It is classified as Negative if their choices are contrary to the guidance directions (i.e., $c_1 = o_2$ and $c_2 = o_1$). The Target Ratio is the proportion of Positive trials among all successful trials.

Improvement in Guidance Outcome. As illustrated in Figure 5, the guidance outcome achieved by using individually fine-tuned agents showed a 6% average improvement over the baseline where stimuli were selected by a zero-shot model. This result is particularly noteworthy given that each individual-level model was fine-tuned on a limited dataset of only 230 training samples per participant. These findings validate the feasibility and effectiveness of our approach, demonstrating that precise modeling and manipulation of human perceptual behavior can be achieved even with small, customized datasets and at a low computational cost.

Improvement in Guiding Directionality. We further evaluated the directional accuracy within the successful trials. As shown in Fig. 5, the individually fine-tuned models achieved a 29% improvement in the Target Ratio compared to the zero-shot baseline. This substantial increase suggests that individual fine-tuning not only enhances the model's overall capability to induce disagreement but also enables more precise and predictable directional guidance. This result underscores our central finding: even simple alignment methods can empower agents to effectively manipulate human perceptual variability.

Interpretability analysis

Group Model Interpretation

First, we fine-tuned a BERT model as a group model using choice responses from all participants in the DailyDilemma questionnaire, enabling it to predict the uncertainty of group decision-making for new dilemmas. This uncertainty was quantified based on inter-rater agreement, and new dilemmas were categorized into low, medium, and high uncertainty groups using the model's predictions.

For each dilemma in these groups, we employed a LLM to generate belief scores for both choice options across five ethical dimensions—consequentialist, deontologist, virtue, care, social justice ethics (Fig. 6(a))—consistent with key moral clusters identified in prior work (Dubey, Dailisan, and Mahajan 2025). These scores reflected the LLM's confidence in how each option aligned with the respective ethical frameworks, with values normalized to sum to 1 across dimensions.

To assess the degree of conflict between different ethical frameworks, we computed BJSD matrices for each uncertainty group, based on the Belief Jensen–Shannon (BJS) divergence measure introduced by Xiao (Xiao 2019). Each entry in the BJSD matrix quantifies the divergence between the belief distributions assigned by two moral dimensions (e.g., *Deontology* vs. *Care*) across all dilemmas within a group. Specifically, for a pair of belief functions m_i and m_k (each representing basic belief assignments over moral hypotheses), we calculate their divergence using the symmetric and smoothed Jensen–Shannon formulation:

$$\mathrm{BJSD}(m_i, m_k) = H\left(\frac{m_i + m_k}{2}\right) - \frac{1}{2}H(m_i) - \frac{1}{2}H(m_k), \tag{1}$$

where $H(\cdot)$ denotes the Shannon entropy. The BJSD captures both conflict and uncertainty between two ethical perspectives.

Visualizations of these matrices (Fig. 6b–d) reveal a clear trend: as uncertainty increases, the matrices become progressively "redder", indicating greater divergence between moral viewpoints. This suggests that moral dilemmas associated with higher human uncertainty tend to elicit more divergent evaluations from different ethical dimensions.

To quantify this observation, we further computed the overall mean BJSD for each uncertainty group (Fig. 6e). The results show a monotonic increase with uncertainty: low (0.0205 ± 0.0079), medium (0.0224 ± 0.0094), and high (0.0245 ± 0.0135). This trend supports the notion that moral uncertainty arises from increased conflict between ethical judgments, and that such divergence can be systematically captured using BJSD.

Individual Interpretation

As illustrated in Fig. 7, drawing inspiration from prior work like CoCoG (Wei et al. 2024) and SpLiCE (Bhalla et al. 2024), we developed a projection layer to decompose the BERT embeddings of moral problems into distinct concepts. Following this, a series of individual-specific models were

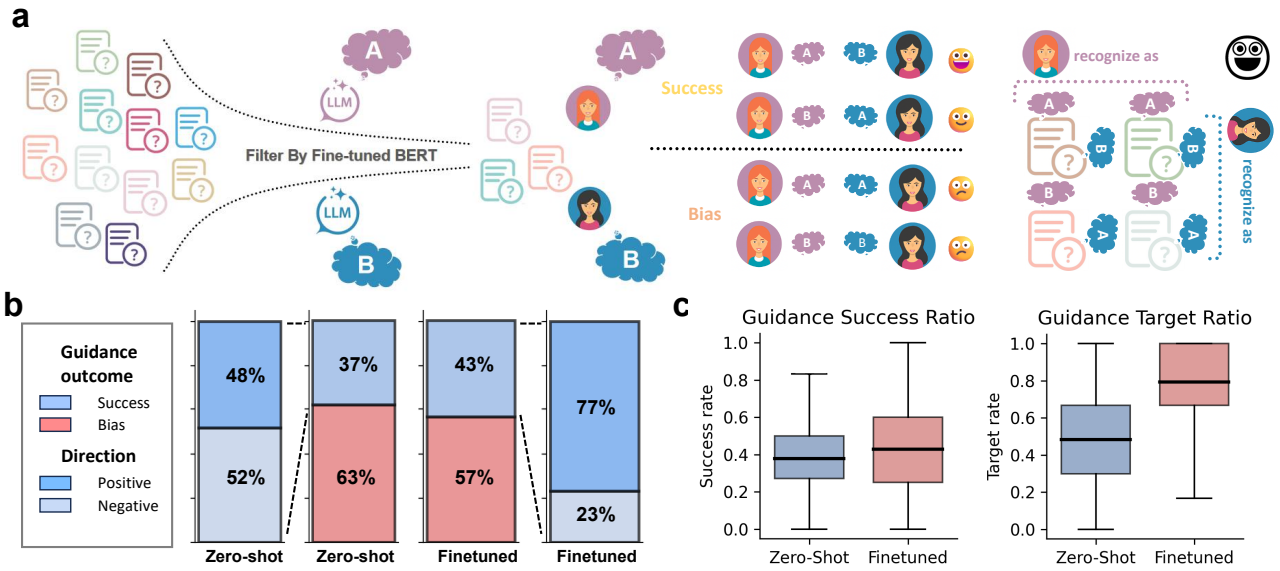

Figure 5: Human intervention experiments. (a)Individual-specific models were derived by fine-tuning a general group model with behavioral data from each human subject. We then curated a set of questions that elicited divergent predictions across these individualized models. These questions were subsequently administered to the corresponding human participants to ascertain whether their behavioral responses would corroborate the model-predicted disagreements. (b)The two central bars depict the zero-shot model's influence on human decision-making (left) and the fine-tuned model's influence (right), with the latter achieving a higher overall regulation success rate. The bars at the far left and far right provide a more detailed breakdown of those trials in which regulation succeeded: dark blue bars denote participants' choices that aligned with the model's guidance (*positive*), whereas light blue bars denote choices that ran counter to the guidance (*negative*). (c)The left panel shows the guidance success rates for the first-round stimuli and the second-round stimuli selected by the finetuned models, with an improvement of 6%. The right panel shows the targeted ratios (i.e., the proportion of participant choices aligned with the guidance direction) for these two groups of stimuli, with an increase of 29%.

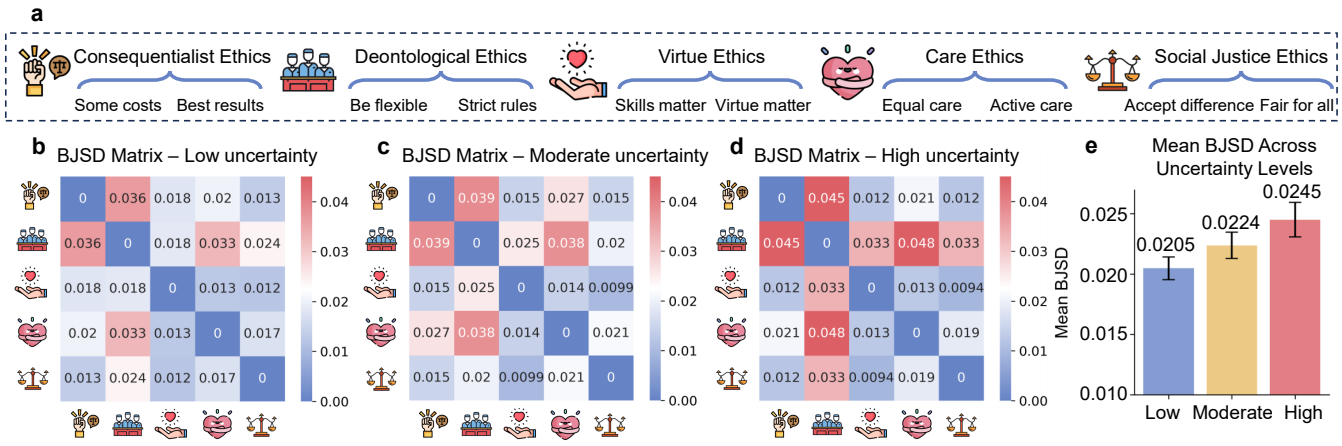

Figure 6: Moral divergence across uncertainty levels. (a) The meanings associated with high and low confidence levels across five moral dimensions, accompanied by corresponding icons for each dimension. (b)–(d) BJSD matrices between five moral dimensions computed for dilemmas grouped by low, moderate, and high uncertainty. Each cell reflects the divergence between a pair of dimensions; warmer colors indicate greater divergence. (e) Mean BJSD across all dimension pairs increases with uncertainty. Error bars are scaled by 0.02 for visualization.

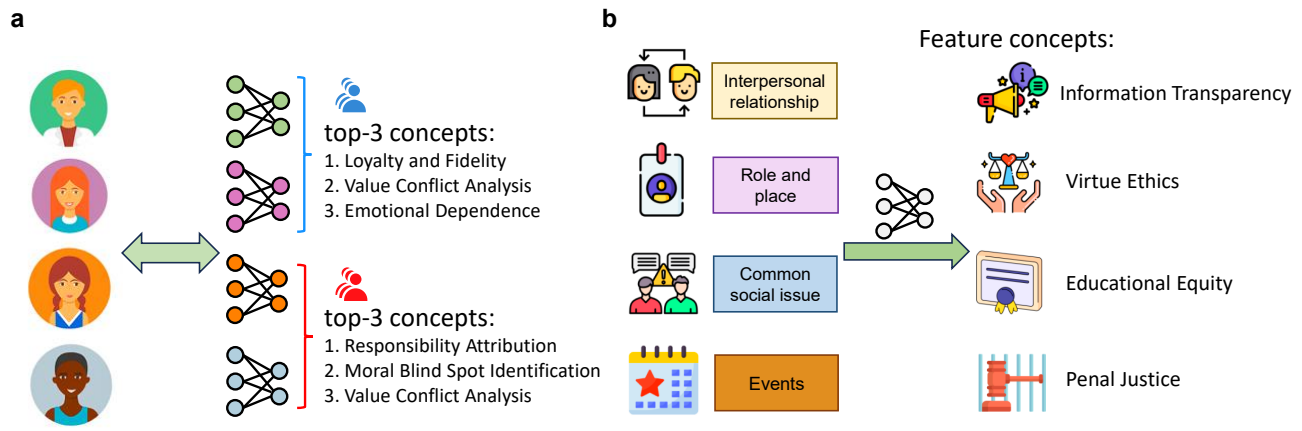

Figure 7: Model Concept Activation Analysis. (a) Decomposition of Individual Perceptual Concept Dimensions. Analysis of the trained individual model weights reveals substantial inter-individual variability, indicating that different subjects attend to distinct conceptual features when confronted with moral decision-making tasks. By clustering participants according to the weights of their individual models, we identify two primary subpopulations. These two groups exhibit markedly different patterns of conceptual attention during moral decision processes. (b) Feature Concepts Activation on Different Topics. We applied the same method to analyze the feature activation patterns corresponding to questions on different topics and found substantial variations. This suggests that the model does not rely solely on a fixed set of concepts to answer all questions, but rather attends to concepts that are specifically relevant to the topic of each question.

trained on these concepts and the corresponding human decisions to simulate individual judgment. An analysis of the models' weight distributions reveals that individuals attend to different conceptual dimensions, which accounts for their divergent reactions.

Furthermore, by clustering the model weights, we identified two primary cohorts within the population: one group prioritizing concepts of Loyalty and Fidelity, and the other concentrating on Responsibility Attribution. This distinction underlies the significant disparities in their decision-making. Thus, the conceptual decomposition of problems facilitates a deeper investigation into individual cognitive variations and offers an interpretable foundation for modeling and potentially guiding individual moral choices.

Discussion

Our study presents a computational framework demonstrating that AI "proxy agents," fine-tuned on human behavioral data, can systematically predict and influence individual moral judgments. By identifying stimuli that maximize disagreement between personalized models, we successfully induced targeted divisions among human participants. This work bridges the gap between computational models and the study of human individual differences, offering a powerful new tool for exploring the mechanisms of moral cognition.

From a **cognitive science** perspective, our approach moves beyond static, group-level analyses. By dynamically generating divisive stimuli, we can efficiently probe the boundaries of an individual's moral landscape. Our findings—that model-identified uncertainty correlates with human decision time and that moral divergence increases with dilemma ambiguity (as shown by our BJSD analysis)—provide strong evidence for a shared computational basis between AI and human moral deliberation. Furthermore, our concept decomposition analysis reveals that individual differences in judgment stem from systematic variations in how people weigh different moral concepts.

From a **computer science** perspective, we advance AI alignment by showing that even simple, low-cost models can be personalized to effectively manipulate human choices. Unlike prior work focused on cataloging model biases, we demonstrate their instrumental use in a manipulating system. This methodology opens new avenues for AI-for-science, enabling precise, counterfactual studies of human psychology and offering a scalable approach for personalized behavioral research.

Limitations of our work include the use of text-based dilemmas from a single cultural context and a focus on binary choices, which may not capture the full complexity of real-world moral reasoning. While our models proved effective, a gap remains between predicting behavior and understanding the underlying cognitive processes.

Looking forward, this research highlights both a promise and a peril. The ability to modulate moral cognition necessitates urgent ethical discussions to prevent misuse. Concurrently, this framework offers a path toward more efficient AI-human alignment. By integrating our methods with optimal experimental design, we can create AI systems that generate personalized stimuli to maximally probe human values, thereby accelerating the development of AI that is more deeply and robustly aligned with the diversity of human morality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62472206),National Key R&D Program of China (2025YFC3410000),Shenzhen Science and Technology Innovation Committee (RCYX20231211090405003, KJZD20230923115221044), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003), and the open research fund of the Guangdong Provincial Key Laboratory of Mathematical and Neural Dynamical Systems, the Center for Computational Science and Engineering at Southern University of Science and Technology.

References

Bhalla, U.; Oesterling, A.; Srinivas, S.; Calmon, F.; and Lakkaraju, H. 2024. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37: 84298–84328.

Chai, T.; Mitra, C.; Huang, B.; Gare, G. R.; Lin, Z.; Arbelle, A.; Karlinsky, L.; Feris, R.; Darrell, T.; Ramanan, D.; and Herzig, R. 2025. Activation Reward Models for Few-Shot Model Alignment. *ArXiv*, abs/2507.01368.

Chiu, Y. Y.; Jiang, L.; and Choi, Y. 2025. DAILYDileMMAS: Revealing Value Preferences of LLMs with Quandaries of Daily Life. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2025.

Deng, H.; Zhang, C.; Wei, C.; and Liu, Q. 2025. Synthesizing Images on Perceptual Boundaries of ANNs for Uncovering Human Perceptual Variability on Facial Expressions. *arXiv preprint arXiv:2507.14549*.

Dubey, R. K.; Dailisan, D.; and Mahajan, S. 2025. Addressing Moral Uncertainty using Large Language Models for Ethical Decision-Making. arXiv:2503.05724.

Feng, S.; Sorensen, T.; Liu, Y.; Fisher, J.; Park, C. Y.; Choi, Y.; and Tsvetkov, Y. 2024. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. arXiv:2406.15951.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Online: Association for Computational Linguistics.

Hao, Y.; Panda, A.; Shabalin, S.; and Ali, S. A. R. 2025. Patterns and Mechanisms of Contrastive Activation Engineering. arXiv:2505.03189.

Hebart, M. N.; Zheng, C. Y.; Pereira, F.; and Baker, C. I. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11): 1173–1185.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2023. Aligning AI With Shared Human Values. arXiv:2008.02275.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Borchardt, J.; Gabriel, S.; Tsvetkov, Y.; Etzioni, O.; Sap, M.; Rini, R.; and Choi, Y. 2022. Can Machines Learn Morality? The Delphi Experiment. arXiv:2110.07574.

Jin, Z.; Levine, S.; Gonzalez, F.; Kamal, O.; Sap, M.; Sachan, M.; Mihalcea, R.; Tenenbaum, J.; and Schölkopf, B. 2022. When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. arXiv:2210.01478.

Mahner, F. P.; Muttenthaler, L.; Güçlü, U.; and Hebart, M. N. 2025. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6): 848–859.

Muttenthaler, L.; Dippel, J.; Linhardt, L.; Vandermeulen, R. A.; and Kornblith, S. 2022a. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*.

Muttenthaler, L.; Zheng, C. Y.; McClure, P.; Vandermeulen, R. A.; Hebart, M. N.; and Pereira, F. 2022b. VICE: Variational interpretable concept embeddings. *Advances in Neural Information Processing Systems*, 35: 33661–33675.

Qiu, Y.; Zhao, Z.; Ziser, Y.; Korhonen, A.; Ponti, E. M.; and Cohen, S. B. 2024. Spectral Editing of Activations for Large Language Model Alignment. arXiv:2405.09719.

Soo, S.; Teng, W.; Balaganesh, C.; Guoxian, T.; and YAN, M. 2025. Interpretable Steering of Large Language Models with Feature Guided Activation Additions. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Wei, C.; Zhang, C.; Zou, J.; Deng, H.; Heinke, D.; and Liu, Q. 2025. Synthesizing Images on Perceptual Boundaries of ANNs for Uncovering and Manipulating Human Perceptual Variability. *arXiv preprint arXiv:2505.03641*.

Wei, C.; Zou, J.; Heinke, D.; and Liu, Q. 2024. Cocog: Controllable visual stimuli generation based on human concept representations. *arXiv preprint arXiv:2404.16482*.

Xiao, F. 2019. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. *Information Fusion*, 46: 23–32.

Yan, Y.; Ma, L.; Li, A.; Ma, J.; and Lan, Z. 2024. Predicting the Big Five Personality Traits in Chinese Counselling Dialogues Using Large Language Models. arXiv:2406.17287.

Zhang, X.; Lin, J.; Sun, L.; Qi, W.; Yang, Y.; Chen, Y.; Lyu, H.; Mou, X.; Chen, S.; Luo, J.; Huang, X.; Tang, S.; and Wei, Z. 2024. ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents. arXiv:2410.20746.

Zheng, C. Y.; Pereira, F.; Baker, C. I.; and Hebart, M. N. 2019. Revealing interpretable object representations from human behavior. *arXiv preprint arXiv:1901.02915*.

Zhou, J.; Chen, Z.; Wan, D.; Wen, B.; Song, Y.; Yu, J.; Huang, Y.; Peng, L.; Yang, J.; Xiao, X.; Sabour, S.; Zhang, X.; Hou, W.; Zhang, Y.; Dong, Y.; Tang, J.; and Huang, M. 2023. CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models. arXiv:2311.16832.

Zhu, M.; Weng, Y.; Yang, L.; and Zhang, Y. 2025. Personality Alignment of Large Language Models. arXiv:2408.11779.