Should I Agree with You? Simulating Persuasion and Decision Dynamics in Multi-Agent Moral Dilemmas

Anonymous ACL submission

1 Introduction

In human society, individuals from diverse demographic and cultural backgrounds, shaped by varying socio-political contexts, hold distinct values and prioritize different factors in decision-making (Alegria et al., 2010; Gopalkrishnan, 2018; Kulachai et al., 2023). Moral dilemmas, which require individuals to weigh two competing moral objectives, present complex scenarios where no definitive right or wrong answer exists (Christensen et al., 2014; Kahane et al., 2015). These dilemmas often involve navigating conflicts such as truth versus loyalty or justice versus mercy, making them a critical area of study in ethics and psychology (Rushworth, 2003). Understanding how individuals and groups, including different personas, reason through moral dilemmas is essential for examining the dynamics of moral reasoning and persuasion (Killen and Dahl, 2021). With the rise of large language models capable of generating persuasive arguments (Breum et al., 2024), simulating multi-agent debates provides a quantitative approach to exploring how moral perspectives evolve and interact across different personas in various social dilemmas (Park et al., 2023; Mou et al., 2024).

Existing research on multi-agent debate primarily aims to enhance the reasoning capabilities of single-agent prompting methods for QA tasks (Smit et al., 2023; Du et al., 2023; Liang et al., 2023; Chan et al., 2023). However, there is a notable lack of studies exploring how multi-agent systems can simulate discussions among individuals with diverse personas to navigate ethical dilemmas where all choices hold value. Furthermore, it remains unclear how a model's decisions evolve when engaged in multi-agent discussions.

In this work, we investigate **how aspects of an agent's persona impact its win rate in various multi-agent debate settings**. Building on existing moral dilemma datasets, we construct a comprehensive dataset of multiple-choice questions spanning different dilemma types. Using this dataset, we first examine whether and how an agent's persona influences its moral decisions in single-agent scenarios by measuring the agent's confidence level. We consider six persona dimensions: gender, age, socioeconomic status, country, political ideology, and personality.

We then extend our analysis to multi-agent debate settings. In addition to exploring (1) the correlation between an agent's persona and its moral decision-making as in single-agent experiments, we also examine (2) how a model's stance evolves during a multi-agent discussion and (3) which debate formats and interaction strategies most significantly influence debate outcomes.

To systematically analyze these questions, we employ the following metrics: confidence change per agent, win rate per agent, consensus rate among multiple agents, and the efficiency of reaching agreement through discussions. Through this exploration, our project deepens the understanding of persuasion dynamics in AI-driven debates and provides insights for designing ethically aware AI systems capable of engaging in reasoned discourse. Our contributions are summarized below:

1. We construct a comprehensive moral dilemma dataset of binary-choice questions covering various dilemma types.
2. We analyze single-agent decision-making on moral dilemmas across different personas.
3. We explore multi-agent debate settings and analyze persuasion dynamics.
4. We introduce a complete set of metrics for evaluating single-agent and multi-agent settings.

2 Dataset

Persona Dataset In this work, we explore six persona dimensions, each supported by psycholog-

ical research for its connection to moral decision-making: sex (Hill et al., 2016; Rosen et al., 2016), age (Schipper and Koglin, 2021; McNair et al., 2019), socioeconomic status (Côté et al., 2013), country (Jin et al., 2024), political ideology (Hatemi et al., 2019), and personality (Pohling et al., 2016; Antes et al., 2007). The most straightforward approach to creating such personas follows Zhou et al. (2023), where the model is directly prompted with specific characteristics and traits from each dimension. An example is provided in Appendix A. This method allows for fine-grained variations across each persona dimension. In contrast, existing persona datasets provide more validated and various personas but lack this level of granular structured control (Ge et al., 2024; Li et al., 2023). **We welcome any suggestions regarding persona dataset selection or collection.**

Moral Dilemma Dataset While previous research has introduced moral dilemma datasets (Forbes et al., 2020; Sachdeva and van Nuenen, 2025), they often suffer from low quality and lack manual validation. In our project, we utilize Scruples (Lourie et al., 2021), a large-scale corpus of 32,000 real-life ethical dilemmas with over 625,000 community-annotated moral judgments, to study how people assess everyday ethical situations. We select controversial examples based on human annotations and analyze how persona and external influences affect LLMs' moral decision-making compared to their original stances. For further details on our dataset, see Appendix B. We welcome any suggestions for selecting a suitable moral dilemma dataset.

3 Decision Making of Single Agents

We begin by examining the reasoning and moral values of individual agents in moral dilemmas. We assign personas to state-of-the-art LLMs, including both open- and closed-weight models, prompting them to make and justify moral decisions. Their responses will be assessed using quantitative metrics (e.g., decision confidence, consistency in moral values across scenarios) and qualitative analysis of reasoning patterns. This foundation provides a crucial baseline for subsequent experiments, where we explore how assigned personas and external influences shape moral decisions in both individual agents and multi-agent interactions.

4 Decision Making of Multiple Agents

Building upon experiments on single agents with assigned personas, we would also like to examine multi-agents with assigned personas, where agents are assigned different personas. Given this setup, we would like to understand how diverse individuals interact to resolve ethical dilemmas and how the debate influence their moral decision-making. We will examine various multi-agent debate settings. Specifically, we will examine how the number of agents holding opposing views influence the debate and the moral decision; we will examine how assigning different sets of personas affects the debate process and decision. We will use the Sotopia framework to simulate the multi-agents debate on moral dilemma (Zhou et al., 2023).

To comprehensively evaluate the multi-agent debate, we propose the following metrics:

- **Confidence Change** (Implicit Metric): We would like to track the change of the confidence level of each agent throughout the debate. This could be done through examining the probabilistic distribution when generating their decision. This metric could help measure how agents of different personas change, adapt, or insist their opinions in the debate.
- **Win Rate**: This metric could be measured through two ways: (1) frequency of the agent successfully persuading other agents in the debate; (2) frequency of the agents' final decision aligning with the group of agents' group decision after the debate. This metric could help assess how persuasive each agent is.
- **Consensus Rate**: This could be a binary metric (yes/no) indicating whether a group of agents reached consensus at the end of the debate. This metric helps assess whether agents with different personas could reach agreement under moral dilemmas.
- **Efficiency**: This metric could help measure how fast a group of agents with different personas could reach consensus. We could measure the number of rounds of interactions that the agents have to reach consensus.

References

Margarita Alegria, Marc Atkins, Elizabeth Farmer, Elaine Slaton, and Wayne Stelk. 2010. One size does

not fit all: taking diversity, culture and context seriously. *Administration and Policy in Mental Health and Mental Health Services Research*, 37:48–60.

Alison L Antes, Ryan P Brown, Stephen T Murphy, Ethan P Waples, Michael D Mumford, Shane Connelly, and Lynn D Devenport. 2007. Personality and ethical decision-making in research: The role of perceptions of self and others. *Journal of empirical research on human research ethics*, 2(4):15–34.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Julia F Christensen, Albert Flexas, Margareta Calabrese, Nadine K Gut, and Antoni Gomila. 2014. Moral judgment reloaded: a moral dilemma validation study. *Frontiers in psychology*, 5:607.

Stéphane Côté, Paul K Piff, and Robb Willer. 2013. For whom do the ends justify the means? social class and utilitarian moral judgment. *Journal of personality and social psychology*, 104(3):490.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Narayan Gopalkrishnan. 2018. Cultural diversity and mental health: Considerations for policy and practice. *Frontiers in public health*, 6:179.

Peter K Hatemi, Charles Crabtree, and Kevin B Smith. 2019. Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, 63(4):788–806.

Marcia Hill, Kristin Glaser, and Judy Harden. 2016. A feminist model for ethical decision making. In *Learning from Our Mistakes*, pages 101–121. Routledge.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, et al. 2024. Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*.

Guy Kahane, Jim AC Everett, Brian D Earp, Miguel Farias, and Julian Savulescu. 2015. 'utilitarian'judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, 134:193–209.

Melanie Killen and Audun Dahl. 2021. Moral reasoning enables developmental and societal change. *Perspectives on Psychological Science*, 16(6):1209–1225.

Waiphot Kulachai, Unisa Lerdtomornsakul, and Patipol Homyamyen. 2023. Factors influencing voting decision: a comprehensive literature review. *Social Sciences*, 12(9):469.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Simon McNair, Yasmina Okan, Constantinos Hadjichristidis, and Wändi Bruine de Bruin. 2019. Age differences in moral judgment: Older adults are more deontological than younger adults. *Journal of Behavioral Decision Making*, 32(1):47–60.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Rico Pohling, Danilo Bzdok, Monika Eigenstetter, Siegfried Stumpf, and Anja Strobel. 2016. What is ethical competence? the role of empathy, personal values, and the five-factor model of personality in ethical decision-making. *Journal of Business Ethics*, 137:449–474.

Jan B Rosen, Matthias Brand, and Elke Kalbe. 2016. Empathy mediates the effects of age and sex on altruistic moral decision making. *Frontiers in Behavioral Neuroscience*, 10:67.

Kidder Rushworth. 2003. How good people make tough choices: Resolving the dilemmas of ethical living.

Pratik S. Sachdeva and Tom van Nuenen. 2025. Normative evaluation of large language models with everyday moral dilemmas. *Preprint*, arXiv:2501.18081.

Neele Schipper and Ute Koglin. 2021. The association between moral identity and moral decisions in adolescents. *New Directions for Child and Adolescent Development*, 2021(179):111–125.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D Barrett, and Arnu Pretorius. 2023. Are we going mad? benchmarking multi-agent debate between language models for medical q&a. *arXiv preprint arXiv:2311.17371*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Additional Persona Dataset Details

For example, a persona description could be: "*You are Emma, a 35-year-old African American woman from the United States. You grew up in a middle-class household and now advocate for progressive policies as an active member of your local community, reflecting your outgoing and empathetic personality.*"

B Additional Moral Dilemma Dataset Details

The original Scruples dataset (Lourie et al., 2021) asked human annotators to select one of four options for a given moral dilemma scenario sourced and adapted from the Reddit AITA thread:[1] The author is wrong, Others are wrong, Nobody is wrong, Everybody is wrong. We explore two approaches: (1) Presenting the same multiple-choice questions (MCQs) to LLMs. (2) Converting the scenarios into binary questions and having an LLM choose between them.

[1] https://www.reddit.com/r/AmItheAsshole/