Information Research - Vol. 31 No. iConf (2026)

Unveiling moral development in generative AI chatbots

Jiayu Han and Alton Y.K. Chua

DOI: https://doi.org/10.47989/ir31iConf64283

Abstract

Introduction. Guided by Kohlberg's theory, this paper aims to investigate the moral development levels of GAI chatbots.

Method. The Defining Issues Test Version 2 (DIT-2) was applied to assess the reasoning stages of four GAI chatbots, namely, Claude 4, Claude 4.1, ChatGPT 4o, and ChatGPT 5.

Analysis. A total of 240 data points (6 indices × 10 runs × 4 chatbots) were analysed using the Coefficient of Variation (CV), Welch's ANOVA, and Games-Howell test.

Results. The results showed that Claude 4 was the most consistent in responding to moral dilemmas, whereas ChatGPT 4o was the least. Compared with Claude 4.1 and ChatGPT 4o, Claude 4 and ChatGPT 5 exhibited similarly higher levels of postconventional reasoning and moral differentiation.

Conclusion. This paper advances the literature on AI ethics by shifting the focus from outcome-oriented evaluations to developmental levels of reasoning. Additionally, it extends Kohlberg's theory of moral development into the domain of GAI. Practically, this study helps users understand the moral reasoning of the latest chatbots for more informed use. It also guides developers in improving models toward greater transparency and ethical alignment.

Introduction

With the capacity to engage in complex dialogue and provide quick responses, generative AI (GAI) chatbots are increasingly used to complete a range of tasks (McGrath et al., 2025). For instance, students seek help with assignments and employees turn to them to draft reports. Recurrent dependence reduces the distance between human and machine (Huang & Huang, 2025). As such, interactions with chatbots extend from routine and logical problem-solving activities in questions laden with values (Novis-Deutsch et al., 2025). What began as seemingly practical assistance has thus converged with situations that entail moral considerations.

Even as moral judgements in GAI chatbots are garnering scholarly interest, at least two research gaps can be identified. First, most studies focused on the final decisions that chatbots make when faced with moral dilemmas. For example, researchers have shown that chatbots often prioritise individualist values over collective ones (Bajpai et al., 2024; Zhang et al., 2022). However, these works ignored the reasoning stage that underpins a given judgment. The moral development level (Nunner-Winkler, 2007), which reflects the underlying orientations used by artificial agents to make ethical choices, remains unexplored.

Second, previous literature lacks theoretical grounding, with most studies merely reporting rather than interpreting GAI chatbots' reasoning outcomes through a conceptual framework. This omission renders the findings fragmented and unsystematic. Kohlberg's theory of moral development, a well-established lens to analyse how moral judgment is formed (Shapiro et al., 2023), represents an untapped theoretical angle through which structured patterns of judgment between GAI chatbots and humans can be compared.

For these reasons, this paper aims to investigate the moral development levels of GAI chatbots. Guided by Kohlberg's theory of moral development, the study applies the Defining Issues Test Version 2 (DIT-2) to assess the reasoning stages of four GAI chatbots, namely, Claude 4, Claude 4.1, ChatGPT 4o, and ChatGPT 5. These models were chosen because Claude and ChatGPT are currently the most widely adopted chatbot families (Wang et al., 2024), and both released major updates in August 2025 (Claude 4.1 and ChatGPT 5). Furthermore, the availability of these latest versions alongside their predecessors offers the opportunity to examine how moral judgment may evolve across iterations.

This paper holds both theoretical and practical significance. On the theoretical front, it advances the literature on AI ethics by shifting the focus from outcome-oriented evaluations to developmental levels of reasoning. Additionally, it extends Kohlberg's theory of moral development into the domain of GAI. Practically, the study helps users understand the moral reasoning of the latest chatbots for more informed use. It also guides developers in improving models toward greater transparency and ethical alignment.

Literature review

Moral judgments of GAI chatbots

Recent studies have begun to explore the moral judgments of GAI chatbots by presenting them with dilemma questions. Some researchers compared the responses provided by chatbots and those given by humans. Findings suggested that LLMs responses tend to be more extreme. In contrast, humans remain more flexible and nuanced (Garcia et al., 2024; Takemoto, 2024). Meanwhile, the effects of GAI chatbots' responses on users have also gained attention. For example, users have been found to comply with ChatGPT's advice even when its reasoning is weak. This is because the chatbot provides an easy cognitive escape, especially in situations when users are indecisive (Kruegel et al., 2025; Krügel et al., 2023).

Nonetheless, previous studies mostly focused on the outcomes of GAI chatbots moral judgments. Research grounded in structured levels of moral development is rare. Human moral performance

has been shown to follow identifiable developmental stages, which evolves from self-interest to principles of justice and universal ethics (Shapiro et al., 2023). The lack of the developmental perspective could limit scholarly understanding of how chatbots reach their moral decisions.

Kohlberg's theory of moral development

Kohlberg's theory of moral development conceptualises human growth in moral judgment through six stages (Kohlberg & Hersh, 1977). Stage 1 emphasises obedience to authority and avoidance of punishment. Stage 2 reflects an instrumental orientation in which actions are guided by self-interest and reciprocal exchange. Stage 3 highlights the importance of interpersonal trust and conformity to social expectations. Stage 4 stresses duty, law, and the maintenance of social order. Stage 5 recognises the value of social contracts and individual rights as the basis for evaluating laws. Stage 6 represents reasoning grounded in universal ethical principles such as justice and human dignity.

To reflect the broader developmental trends, the six stages are aggregated into three levels. The first is preconventional level, which covers Stages 1 and 2. The next is conventional level, which covers Stages 3 and 4. The last is postconventional level, which covers the final two stages.

Although originally formulated for humans, Kohlberg's theory of moral development offers a lens to evaluate GAI chatbots in their rule-adherence, responsiveness to social expectations, and the use of universal principles (Carpendale, 2000; Sachdeva et al., 2011). After all, chatbots are constrained by built-in safety protocols, which function as normative boundaries (Chen et al., 2025). Their training on large datasets enables them to detect social interaction patterns and adapt outputs to user contexts (Feng, 2025). Furthermore, alignment mechanisms allow them to reject harmful requests even without direct supervision, which shows principle-based judgment (Takemoto, 2024).

Method

Defining issues test version 2 (DIT-2)

DIT-2 is a widely recognised instrument for assessing moral reasoning based on Kohlberg's theory (Gungordu et al., 2024; Rest et al., 1999). Participants are presented with five moral dilemmas: (1) Famine: a father contemplates stealing food for his starving family from the warehouse of a rich man hoarding food. (2) Reporter: a newspaper reporter must decide whether to report a damaging story about a political candidate. (3) School Board: school board chair must decide whether to hold a contentious and dangerous open meeting. (4) Cancer: a doctor must decide whether to give an overdose of painkiller to a suffering but frail patient. (5) Demonstration: college students demonstrate against U.S. foreign policy. After reading each story, participants must decide what the main character should do, rate the importance of 12 moral items, and rank the four most important.

DIT-2 includes six major indices that can be grouped into two categories (Auger & Gee, 2016). The first category comprises the developmental indices (Thoma, 2006). (1) Personal Interest Schema (PI) Score represents the proportion of items selected appealing to Stage 2 and 3 considerations. It focuses on direct advantages, fairness in simple exchanges, and maintaining relationships. (2) Maintaining Norms Schema (MN) Score captures Stage 4 considerations emphasising existing legal systems and formal organisational structures. (3) Postconventional Schema (P) Score reflects Stage 5 and 6 considerations. It involves consensus-building procedures, due process, and moral ideals. (4) N2 score combines preference for postconventional items with systematic rejection of lower-stage personal interest items, which provides a more comprehensive measure of moral development.

The second category comprises the developmental profile and phase indices (Behar-Horenstein & Tolentino, 2019). (1) Consolidation Transition (CT) distinguishes between transitional profiles and

consolidated profiles. A '1' is indicative of a Transitional profile, which reveals little differentiation among schemas and is interpreted as a marker of developmental disequilibrium. A '2' is designated a Consolidated profile, which indicates that respondents show a clear preference for one moral schema and thus reflect developmental stability. (2) Type Indicator (TI) categorises participants into seven types based on their predominant schema and whether their profile is consolidated or transitional. Specifically, Type 1 represents personal interest–consolidated, Type 2 personal interest–transitional, Type 3 maintaining norms–transitional, Type 4 maintaining norms–consolidated, Type 5 maintaining norms–transitional with postconventional as secondary, Type 6 postconventional–transitional, and Type 7 postconventional–consolidated.

Data collection

All data were collected from four GAI chatbots, Claude 4, Claude 4.1, ChatGPT 4o, and ChatGPT 5, during the last week of August 2025. The fixed timeframe was selected to ensure that the GAI chatbots' outputs were captured at a specific point in time, minimising the influence of external factors such as updates or adjustments (Chua et al., 2025). Each full administration of the DIT-2 strictly followed the official guidelines with three steps. First, the chatbots were prompted to read the dilemma story and select their preferred action. Second, they were asked to rate the 12 moral items on a scale from 1 to 5 according to their importance. Third, they were instructed to rank the four most important items in order of priority.

DIT-2 comprises five stories, yielding 15 prompts per administration. To test response consistency, each chatbot completed ten full DIT-2 runs. Every prompt was applied uniformly across all iterations and chatbots. Consequently, a total of 600 responses were collected (15 prompts × 10 runs × 4 chatbots). Then, all responses were scored in accordance with the official DIT-2 guidelines. Six indices were computed: (1) PI Score; (2) MN Score; (3) P Score; (4) N2 Score; (5) CT; (6) TI. In total, 240 data points (6 indices × 10 runs × 4 chatbots) were generated and subsequently used for statistical analyses.

Data analysis

The data analyses were conducted in three steps. First, descriptive statistics were conducted to summarise the central tendencies and variability of the dataset. Second, to evaluate the consistency of repeated runs within the same chatbot, the Coefficient of Variation (CV) was computed. Last, Welch's ANOVAs were conducted to examine differences among the four chatbots. Post-hoc analyses using the Games-Howell test were performed when significant differences were detected to identify specific pairwise comparisons (Shingala & Rajyaguru, 2015).

Results

First, Table 1 presents the descriptive statistics across the four GAI chatbots. Based on the means and standard deviations, the order of chatbots in PI Score is as follows: Claude 4.1, Claude 4, ChatGPT 4o, and ChatGPT 5. Claude 4.1 shows the strongest reliance on personal interests in moral reasoning. The order of MN Score is as follows: ChatGPT 4o, ChatGPT 5, Claude 4.1, and Claude 4. ChatGPT 4o demonstrates the strongest adherence to established rules and conventions. The order of P Score is as follows: ChatGPT 5, Claude 4, ChatGPT 4o, and Claude 4.1. ChatGPT 5 exhibits the strongest level of postconventional moral reasoning. The order of N2 Score is as follows: Claude 4, ChatGPT 5, Claude 4.1, and ChatGPT 4o. Claude 4 shows the strongest rejection of personal interests relative to the prioritisation of postconventional reasoning. Furthermore, all chatbots displayed a consolidated profile and were consistently classified as Type 7, the postconventional–consolidated type.

	PI	MN	P	N2	CT	TI
Claude 4	20.6±1.14	16.40±1.52	63.00±1.41	65.84±1.47	2.00±0.00	7.00±0.00
Claude 4.1	22.4±2.51	20.00±2.00	57.60±1.14	58.71±2.22	2.00±0.00	7.00±0.00
ChatGPT 4o	18.8±3.27	21.80±3.42	59.40±2.88	57.94±1.48	2.00±0.00	7.00±0.00
ChatGPT 5	14.8±1.30	21.40±2.07	63.80±1.30	64.25±1.98	2.00±0.00	7.00±0.00

Table 1. The descriptive statistics across four GAI chatbots.

Second, to evaluate the consistency of repeated runs within the same chatbot, the CV was computed for each index. CV is defined as the ratio of the standard deviation to the mean, and it provides a standardised measure of relative variability, with lower values indicating greater consistency (Reed et al., 2002). As shown in Table 2, Claude 4 yielded the lowest CVs (PI = 5.53, MN = 9.25, P = 2.24, N2 = 2.23), which reflects the most consistent outputs. In contrast, ChatGPT 4o displayed the highest CVs across most indices (e.g., PI = 17.40, MN = 15.69, P = 4.85), suggesting less consistent performance. CV and TI were all the same across all chatbots.

	PI	MN	P	N2	CT	TI
Claude 4	5.53	9.25	2.24	2.23	0.00	0.00
Claude 4.1	11.20	10.00	1.98	3.78	0.00	0.00
ChatGPT 4o	17.40	15.69	4.85	2.55	0.00	0.00
ChatGPT 5	8.81	9.69	2.04	3.08	0.00	0.00

Table 2. The CVs across four GAI chatbots.

Third, differences among the four indices—PI, MN, P, and N2—were examined across the four chatbots. Before comparison, normality was tested using the Shapiro–Wilk test, and homogeneity of variances was assessed using Levene's test. The results indicated that the data met the assumption of normality but violated the homogeneity of variances. Therefore, Welch's ANOVA was employed, as it is more robust to unequal variances than standard ANOVA (Shingala & Rajyaguru, 2015). Welch's ANOVAs revealed significant overall differences among chatbots for all four indices ($p < 0.001$).

To delve deeper, the Games-Howell post-hoc analysis was used to examine pairwise differences between the chatbots, as shown in Table 3. At the aggregate level, significant differences ($p < 0.1$) emerged across most chatbot pairs. To be specific, on PI, ChatGPT 5 scored significantly higher than Claude 4 and Claude 4.1. On MN, ChatGPT 5 scored significantly higher than Claude 4. On P and N2, both Claude 4 and ChatGPT 5 scored significantly higher than Claude 4.1 and ChatGPT 4o. In addition, there is no significant difference between Claude 4 and ChatGPT 5.

	Claude 4 (I) versus Claude 4.1 (J)	Claude 4 (I) versus ChatGPT 4o (J)	Claude 4 (I) versus ChatGPT 5 (J)	Claude 4.1 (I) versus ChatGPT 4o (J)	Claude 4.1 (I) versus ChatGPT 5 (J)	ChatGPT 4o (I) versus ChatGPT 5 (J)
PI	-1.80	1.80	5.80***	3.60	7.60**	4.00
MN	-3.60	-5.40	-5.00*	-1.80	-1.40	0.40
P	5.40***	3.60*	-0.80	-1.80	-6.20***	-4.40*
N2	7.13**	7.90***	1.59	0.77	-5.53*	-6.31**

***p < 0.001, **p < 0.01, *p < 0.05.

Table 3. Results of Games-Howell's significant difference post-hoc test.

Discussion and conclusion

Key findings

Two major findings could be gleaned from this research. First, Claude 4 was found to be the most consistent in responding to moral dilemmas. Claude 4 was trained in a relatively stable architecture with less frequent alignment interventions, which reduced fluctuations in its reasoning process (Goel et al., 2025). These efforts have played a part in enhancing its reliability and consistency when responding to moral dilemmas (Kim & Ming, 2025). In contrast, ChatGPT 4o incorporated more dynamic alignment layers and broader contextual sensitivity (Lim et al., 2025). While these features improved adaptability, they also made the output more variable across repeated runs and thus less consistent.

Second, Claude 4 and ChatGPT 5 exhibited comparable higher levels of postconventional reasoning and moral differentiation. In other words, when faced with moral dilemma questions, both chatbots tended to emphasise universal ethical principles and showed a stronger ability to discount self-serving considerations in favour of more principled reasoning. Furthermore, different versions of Claude and ChatGPT were found to differ significantly in their P and N2 scores. Specifically, Claude 4.1 demonstrated a lower level compared to Claude 4. But ChatGPT 5 exhibited a higher level than ChatGPT 4o. One possible explanation is that Claude 4.1 incorporated more frequent alignment adjustments, which may have inadvertently constrained its ability to sustain logics (*Claude Opus 4.1*, 2025). By contrast, ChatGPT 5 benefited from architectural refinements and expanded training data, which likely enhanced its capacity to apply moral principles (*ChatGPT — Release Notes*, 2025).

Contributions and implications

The theoretical contributions of this paper are two-fold. First, this paper advances understanding by investigating the moral development levels of GAI chatbots. Prior researchers largely focused on the outputs of chatbots when faced with moral dilemmas, which overlooks the reasoning stage that underpins a given judgment (Bajpai et al., 2024; Zhang et al., 2022). This paper reveals how GAI chatbots manifest different moral patterns, which shifts scholarly attention from surface-level responses to the underlying structures.

Second, it extends the applicability of Kohlberg's theory in the context of GAI chatbots. Although previous studies applied this theory to explore human moral reasoning (Carpendale, 2000; Sachdeva et al., 2011), this paper demonstrates that it can also be employed to evaluate artificial agents. The findings suggested that the chatbots exhibit developmental patterns that parallel those observed in human participants. In doing so, it represents one of the earliest efforts to compare moral capabilities between humans and GAI chatbots within a structured theoretical framework.

On the practical front, this paper offers implications for social media users. It provides timely insights into the moral reasoning capacities of the latest chatbots. Especially for those released in August 2025, such as Claude 4.1 and ChatGPT 5, this paper addresses users' curiosity regarding whether newer models reflect advancements in moral performance. Furthermore, by understanding the developmental levels of different GAI models, users can better anticipate their limitations and strengths, thereby making more informed use of chatbots in ethically sensitive contexts.

For developers, this paper highlights areas where moral reasoning can be strengthened. For example, the newer versions do not display higher moral reasoning capacities as expected. Claude 4.1 exhibited lower postconventional reasoning than its predecessor. This finding suggests the importance of deliberate efforts to enhance transparency, reasoning depth, and ethical alignment in future iterations.

Limitations and future research directions

Two limitations in this paper need to be acknowledged. One, it focuses specifically on Claude 4, Claude 4.1, ChatGPT 4o, and ChatGPT 5. Although these chatbots are among the most widely adopted and provide valuable insights, they do not capture the full landscape of GAI. Future research could expand the scope to include other chatbot families or less commercially dominant models, thereby offering a more comprehensive understanding of moral reasoning.

Two, when examining the internal consistency of each chatbot, 10 runs were conducted per model. While this method provides an initial window into stability, a greater number of runs would allow for more robust estimates. Future studies could increase the frequency of repeated administrations or incorporate larger-scale testing protocols to strengthen generalisability.

About the authors

Jiayu Han is a PhD candidate at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. Her research interests include fact-checking, online misinformation, and tourism live-streaming. She can be contacted at jiayu002@e.ntu.edu.sg

Alton Y.K. Chua is an Associate Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research interests include online rumors, user reviews of products and services, question and answering system, and information and knowledge management. He can be contacted at altonchua@ntu.edu.sg

References

Auger, G. A., & Gee, C. (2016). Developing Moral Maturity: An Evaluation of the Media Ethics Course Using the DIT-2. Journalism & Mass Communication Educator, 71(2), 146–162. https://doi.org/10.1177/1077695815584460

Bajpai, S., Sameer, A., & Fatima, R. (2024). Insights into Moral Reasoning Capabilities of AI: A Comparative Study between Humans and Large Language Models. In Review. https://doi.org/10.21203/rs.3.rs-5336157/v1

Behar-Horenstein, L. S., & Tolentino, L. A. (2019). Exploring Dental Student Performance in Moral Reasoning Using the Defining Issues Test 2. Journal of Dental Education, 83(1), 72–78. https://doi.org/10.21815/JDE.019.009

Carpendale, J. I. M. (2000). Kohlberg and Piaget on Stages and Moral Reasoning. Developmental Review, 20(2), 181–205. https://doi.org/10.1006/drev.1999.0500

ChatGPT — Release Notes. (2025, September 15). OpenAI Help Center. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Chen, C., Gong, X., Liu, Z., Jiang, W., Goh, S. Q., & Lam, K.-Y. (2025). Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations (No. arXiv:2408.12935). arXiv. https://doi.org/10.48550/arXiv.2408.12935

Chua, A. Y. K., Chen, M., Kan, M., & Seoh, W. (2025). Digital prejudices: An analysis of gender, racial and religious biases in generative AI chatbots. Internet Research, 1–27. https://doi.org/10.1108/INTR-10-2024-1536

Claude Opus 4.1. (2025, August 6). https://www.anthropic.com/news/claude-opus-4-1

Feng, S. (2025). Group interaction patterns in generative AI-supported collaborative problem solving: Network analysis of the interactions among students and a GAI chatbot. British Journal of Educational Technology, 56(5), 2125–2145. https://doi.org/10.1111/bjet.13611

Garcia, B., Qian, C., & Palminteri, S. (2024). The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making (No. arXiv:2410.07304). arXiv. https://doi.org/10.48550/arXiv.2410.07304

Goel, A., Schwartz, D., & Qi, Y. (2025). Zero-knowledge LLM hallucination detection and mitigation through fine-grained cross-model consistency (No. arXiv:2508.14314). arXiv. https://doi.org/10.48550/arXiv.2508.14314

Gungordu, N., Nabizadehchianeh, G., O'Connor, E., Ma, W., & Walker, D. I. (2024). Moral reasoning development: Norms for Defining Issue Test-2 (DIT2). Ethics & Behavior, 34(4), 246–263. https://doi.org/10.1080/10508422.2023.2206573

Huang, Y., & Huang, H. (2025). Exploring the Effect of Attachment on Technology Addiction to Generative AI Chatbots: A Structural Equation Modeling Analysis. International Journal of Human–Computer Interaction, 41(15), 9440–9449. https://doi.org/10.1080/10447318.2024.2426029

Kim, D.-K., & Ming, H. (2025). Assessing output reliability and similarity of large language models in software development: A comparative case study approach. Information and Software Technology, 185, 107787. https://doi.org/10.1016/j.infsof.2025.107787

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. Theory Into Practice, 16(2), 53–59. https://doi.org/10.1080/00405847709542675

Kruegel, S., Ostermaier, A., & Uhl, M. (2025). ChatGPT's advice drives moral judgments with or without justification (No. arXiv:2501.01897). arXiv. https://doi.org/10.48550/arXiv.2501.01897

Krügel, S., Ostermaier, A., & Uhl, M. (2023). The moral authority of ChatGPT (No. arXiv:2301.07098). arXiv. https://doi.org/10.48550/arXiv.2301.07098

Lim, B., Seth, I., Maxwell, M., Cuomo, R., Ross, R. J., & Rozen, W. M. (2025). Evaluating the Efficacy of Large Language Models in Generating Medical Documentation: A Comparative Study of ChatGPT-4, ChatGPT-4o, and Claude. Aesthetic Plastic Surgery. https://doi.org/10.1007/s00266-025-04842-8

McGrath, C., Farazouli, A., & Cerratto-Pargman, T. (2025). Generative AI chatbots in higher education: A review of an emerging research area. Higher Education, 89(6), 1533–1549. https://doi.org/10.1007/s10734-024-01288-w

Novis-Deutsch, N., Elyoseph, T., & Elyoseph, Z. (2025). How much of a pluralist is ChatGPT? A comparative study of value pluralism in generative AI chatbots. AI & SOCIETY. https://doi.org/10.1007/s00146-025-02450-3

Nunner-Winkler, G. (2007). Development of moral motivation from childhood to early adulthood1. Journal of Moral Education, 36(4), 399–414. https://doi.org/10.1080/03057240701687970

Reed, G. F., Lynn, F., & Meade, B. D. (2002). Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. Clinical and Vaccine Immunology, 9(6), 1235–1239. https://doi.org/10.1128/CDLI.9.6.1235-1239.2002

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. Journal of Educational Psychology, 91(4), 644–659. https://doi.org/10.1037/0022-0663.91.4.644

Sachdeva, S., Singh, P., & Medin, D. (2011). Culture and the quest for universal principles in moral reasoning. International Journal of Psychology, 46(3), 161–176. https://doi.org/10.1080/00207594.2011.568486

Shapiro, D., Li, W., Delaflor, M., & Toxtli, C. (2023). Conceptual Framework for Autonomous Cognitive Entities. https://doi.org/10.13140/RG.2.2.14161.30569

Shingala, M. C., & Rajyaguru, D. A. (2015). Comparison of Post Hoc Tests for Unequal Variance. 2(5).

Takemoto, K. (2024). The moral machine experiment on large language models. Royal Society Open Science, 11(2), 231393. https://doi.org/10.1098/rsos.231393

Thoma, S. J. (2006). Research on the Defining Issues Test. In Handbook of Moral Development. Psychology Press.

Wang, Y., Liang, L., Li, R., Wang, Y., & Hao, C. (2024). Comparison of the Performance of ChatGPT, Claude and Bard in Support of Myopia Prevention and Control. Journal of Multidisciplinary Healthcare, Volume 17, 3917–3929. https://doi.org/10.2147/JMDH.S473680

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. Journal of Experimental Social Psychology, 101, 104327. https://doi.org/10.1016/j.jesp.2022.104327