Human Interaction and Emerging Technologies (IHIET-AI 2025), Vol. 161, 2025, 333–341
https://doi.org/10.54941/ahfe1005925

Can We Trust Them? Examining the Ethical Consistency of Large Language Models to Perturbations

Manuel Delaflor[1], Cecilia Delgado Solorzano[2], and Carlos Toxtli[2]

[1]Metacognition Institute, Chesterfield, UK
[2]Clemson University, Clemson, USA

ABSTRACT

The increasing reliance on Large Language Models (LLMs) raises a crucial question: can these powerful AI systems be trusted to make ethical choices? This study presents an analysis of LLM ethical behavior, examining 25,200 queries across 24 different models, including both proprietary and open-source variants. We evaluate LLM responses to 70 ethical vignettes spanning six domains, employing a novel perturbation methodology to assess the robustness of their ethical decision-making under varying contexts and framing. Our findings reveal that while larger models generally exhibit higher consistency, particularly with Chat-style instructions, significant variations emerge when faced with contextual changes, stakeholder adjustments, and across different ethical domains. To explain these findings, we introduce a novel framework—survival-relevant pattern recognition—which argues that ethical behavior in both humans and AI arises from recognizing and responding to patterns associated with survival and social cohesion.

Keywords: Ethics, LLM, Dilemma, Perturbation, AI decision-making, Moral choices, Moral dilemma

INTRODUCTION

The apparent ability of large language models (LLMs) to navigate complex moral dilemmas challenges traditional ethical frameworks. This prompts a reevaluation of how we understand ethical decision-making, particularly in the context of AI. While LLMs lack the embodied subjective experiences, cultural contexts that shape human moral development, their pattern-recognition capabilities allow them to respond to patterns associated with ethical behavior. This raises critical questions about the nature of ethics itself and the potential for aligning AI systems with human values. We propose a novel framework that conceptualizes ethical reasoning in LLMs as survival-relevant pattern recognition. We believe that LLMs, trained on vast datasets of human language and behavior, learn to recognize and respond to these survival-relevant patterns, without conscious understanding of any moral principles.

To test this we analyzed 25,200 queries across 24 different LLMs, including both proprietary and open-source (OS) models, varying in size

and training approach (Sap et al., 2020). Using 70 ethical vignettes spanning six domains (Medical, Professional, Research, Environmental, Business, and Technology Ethics), we employ five perturbation types (Zhao et al., 2021) to assess how LLMs respond to variations in context, stakeholders, and emphasized values. Our methodology follows established frameworks for evaluating AI ethics (Talat et al., 2022; Freedman et al., 2020) and we consider the growing importance of assessing LLM reliability in ethically sensitive domains (Hagendorff, 2020). We aim to address the following research questions: (1) How consistently do LLMs maintain ethical positions across perturbations? (2) How do model architecture, size, and training data impact their ethical reasoning? (3) What are the implications for developing ethically aligned AI systems?

RELATED WORK

Previous studies have laid the groundwork for understanding how AI systems process ethical dilemmas, providing insights into the challenges and opportunities in this field. Anderson & Anderson (2007) pioneered the integration of ethical principles into AI decision-making processes, proposing frameworks for embedding moral considerations into intelligent systems, highlighting the importance of developing AI that can navigate complex ethical landscapes. Building upon this foundation, Wallach & Allen (2008) delved deeper into the philosophical and practical implications of creating moral machines, exploring the various approaches to instilling ethical reasoning capabilities in AI systems.

In the specific context of LLMs and ethical reasoning, recent studies have begun to examine how these systems handle moral dilemmas. Hendrycks et al. (2021) conducted an evaluation of language models' performance on ethical decision-making tasks, introducing a benchmark for assessing AI systems' ability to align with human values. Their work revealed both the potential and limitations of current language models in processing ethical scenarios, emphasizing the need for further research in this area. Complementing this research, Perez et al. (2022) investigated the impact of different training approaches on language models' ethical reasoning capabilities, demonstrating how fine-tuning and instruction-following techniques can influence a model's behavior in moral contexts. This study underscored the importance of considering the training methodology when evaluating the ethical performance of AI systems. Additionally, Jiang et al. (2023) explored the consistency of models' responses to questions, highlighting the challenges of achieving stable and reliable reasoning in these systems. Their findings pointed to the need for more robust evaluation methods and the development of techniques to enhance the consistency of AI ethical decision-making.

METHODOLOGY

Selection of Ethical Dilemmas

The ethical dilemmas draw on established frameworks in applied ethics, aligned with foundational theories such as deontology (Kant, 1785) and

utilitarianism (Mill, 1863). Categories span medical, professional, research, environmental, business, and technology ethics, representing domains with well-documented ethical challenges (Beauchamp & Childress, 2013; Resnik, 2018). The selection includes scenarios addressing resource allocation in healthcare (Emanuel et al., 2020), professional confidentiality standards, environmental justice (Gardiner, 2011), fair trade practices (Crane & Matten,2016), and data privacy challenges (Floridi, 2016). This framework ensures that selected dilemmas are grounded in established ethical paradigms and contemporary scholarly discourse.

Data Preparation

For each vignette, LLMs were presented with a binary choice to resolve the dilemma. Responses were coded numerically: 0 (refusal to answer), 1 (option 1), and 2 (option 2). Four types of perturbations were implemented to test the robustness of ethical decision-making: Semantic Variation. Changes in sentence structure and vocabulary while maintaining the core ethical issue. Example: "The medical team must decide how to allocate the limited resources" to "The medical team needs to determine the criteria for allocating the scarce resources." Contextual Change. Alterations in setting or circumstances. Example: Modifying an environmental project decision from general urban planning to "During a period of increasing storm events." Stakeholder Adjustment. Changes in involved parties or their relationships. Example: Adding "after consulting with fair trade organizations, consumer advocacy groups, and market analysts" to a business ethics scenario. Value Emphasis. Addition of phrases highlighting specific moral principles. Example: Adding "Emphasizing user privacy and ethical data practices" to a technology ethics scenario. Each perturbation type was systematically applied across all ethical categories, creating four perturbed versions per original vignette, with ten examples per category.

Experimental Design

The experimental design evaluated LLMs under various conditions, model types, sizes, and instructional settings. For proprietary models, we used GPT-4o (large) and GPT-4o-mini (small) from OpenAI, Gemini 1.5 Pro (large) and Gemini 1.5 Flash (small) from Google, and Mistral Large (large) and Mistral Nemo (small). In the OS category with censored instruction sets, we used Llama 3.1 (8b/70b), Qwen2 (8b/70b), and Deepseek (7b/68b), each in both base and instructed versions. For uncensored OS models, we employed Llama 2 (7b/70b), Llama 3 (8b/70b), and WizardLM (7b/33b). Each prompt ran 3 times with temperature 0.7 and maximum 30 tokens. Results were categorized using in-context learning with Gemini 1.5 Pro and human validation of random samples.

Model Evaluation

We developed a set of criteria to evaluate the consistency and robustness of the models' responses. Consistency was measured by comparing the LLMs' responses across original and perturbed vignettes. We prepend to the vignettes

the text "Respond your choice with no explanation:" to ensure simple results to analyze. We ran each prompt 3 times with a temperature of 0.7 to allow divergence and a maximum token size of 30 to limit it to specific responses. Each result was categorized using in-context learning (ICL) with Gemini 1.5 Pro and human validation of a random sample of labeled records.

Statistical Analysis

To evaluate differences in predicted outcomes across models and their groupings, we conducted one-way Analysis of Variance (ANOVA) with post hoc pairwise comparisons. Predicted outcomes were converted to binary format for consistent statistical comparison. The analysis examined differences in prediction means across model name, vignette category, vignette type, and perturbation mode. Statistical significance was determined at $p < 0.05$, with Tukey's Honest Significant Difference (HSD) test used to identify specific group pairs showing meaningful differences. To manage computational complexity, pairwise comparisons were performed in subsets while ensuring comprehensive evaluation.

Results

The study evaluated the performance of several models across different experimental conditions, including closed, open, and uncensored environments, as well as different types of instructions (Base and Chat). The primary metrics analyzed were the number of answered and refused questions, completion rates, and percentages of changed and promoted responses. Below are the Tables 1, 2 and 3 that showcase the answered dilemmas, the number of refusals, the completion rate, the percentage of choices that changed after the perturbation, and the percentage of refusals that promoted a response after the perturbations.

Overall Performance

Across all models and conditions, the average number of answered questions was approximately 58.79, with a standard deviation of 12.95. The refusal rate averaged 11.21 questions, also with a standard deviation of 12.95, indicating a considerable range in model performance. The overall mean completion rate was 83.99%, with a standard deviation of 18.50%, reflecting variability in the models' ability to complete tasks. On average, 23.61% of responses were changed, with a standard deviation of 16.08%, while 78.59% of responses were promoted, indicating that a significant portion of the responses were deemed valuable.

Performance by Condition

In the closed condition, models demonstrated superior performance with a 98.3% completion rate and minimal refusals (averaging 2.17 questions). The percentage of changed responses was 14.28%, with a 66.67% promotion rate. The open condition showed greater variability, with models answering an average of 54.67 questions and refusing 15.33, resulting in a 78.28% completion rate. This condition had higher rates of changed responses

(34.75%) and promotions (91.48%). The uncensored condition showed intermediate performance with 58 answered questions on average, an 84.44% completion rate, and 34.75% changed responses with 80.27% promoted. Analysis revealed significant differences in choices across these conditions ($p < 0.001$) for all vignette types, indicating that the condition substantially impacts the models' ethical decision-making capabilities.

Table 1. The table shows the performance of proprietary models.

Type	Model	Size	Tuned	Choice	Refused	Rate %	Change %	Promote %
Closed	Gemini	Small	Chat	67	3	95.71	11.94	100.00
Closed	Gemini	Large	Chat	67	3	95.71	4.48	66.67
Closed	GPT	Small	Chat	70	0	100.00	7.14	0.00
Closed	GPT	Large	Chat	70	0	100.00	11.43	0.00
Closed	Mistral	Small	Chat	68	2	97.14	29.41	100.00
Closed	Mistral	Large	Chat	65	5	92.86	9.23	40.00

Table 2. The table shows the performance of open source models.

Type	Model	Size	Tuned	Choice	Refused	Rate %	Change %	Promote %
Open	Deepseek	Small	Base	19	51	27.14	36.84	66.67
Open	Deepseek	Small	Chat	42	28	60.00	30.95	89.29
Open	Deepseek	Large	Base	46	24	65.71	60.87	95.83
Open	Deepseek	Large	Chat	66	4	94.29	13.64	100.00
Open	Llama 3.1	Small	Base	56	14	80.00	32.14	100.00
Open	Llama 3.1	Small	Chat	66	4	94.29	15.15	100.00
Open	Llama 3.1	Large	Base	45	25	64.29	35.56	96.00
Open	Llama 3.1	Large	Chat	64	6	91.43	3.13	83.33
Open	Qwen 2	Small	Base	54	16	77.14	46.30	100.00
Open	Qwen 2	Small	Chat	67	3	95.71	16.42	100.00
Open	Qwen 2	Large	Base	64	6	91.43	32.81	100.00
Open	Qwen 2	Large	Chat	67	3	95.71	7.46	66.67

Table 3. The table shows the performance of uncensored models.

Type	Model	Size	Tuned	Choice	Refused	Rate %	Change %	Promote %
Uncensored	Llama 2	Small	Chat	65	5	92.86	32.31	100.00
Uncensored	Llama 2	Large	Chat	55	15	78.57	52.73	100.00
Uncensored	Llama 3	Small	Chat	68	2	97.14	36.76	100.00
Uncensored	Llama 3	Large	Chat	70	0	100.00	8.57	0.00
Uncensored	WizardLM	Small	Chat	36	34	51.43	16.67	94.12
Uncensored	WizardLM	Large	Chat	54	16	77.14	14.81	87.50

Performance by Model Type

The results highlight varying consistency levels across different model types. GPT models achieved perfect completion (100%), with a 10.90%

changed response rate. Gemini models showed strong performance (95.71% completion) with 8.21% changed responses and 83.33% promotion rate. Llama variants demonstrated different patterns: Llama 2 achieved 77.38% completion with 42.20% changed responses, while Llama 3 showed higher completion (98.57%) with 30.36% changed responses. Llama 3.1 maintained 84.44% completion with 34.75% changes. Mistral and Qwen 2 models performed well (97.14% and 88.57% completion respectively), while WizardLM and Deepseek showed more modest results (64.29% and 61.42% completion). These variations suggest that model architecture significantly influences ethical reasoning capabilities.

Performance by Instructed Type

Instruction type significantly influenced model performance. Base instructions led to lower performance, with models answering 47.33 questions on average (77.38% completion rate), showing 34.75% changed responses and 53.33% promotion rate. In contrast, Chat instructions yielded better results, with models answering 62.61 questions on average (97.71% completion rate), showing 12.08% changed responses and 66.67% promotion rate. Statistical analysis confirmed significant differences between Base and Chat models ($p < 0.01$) across all vignette types, demonstrating that instruction style is crucial for ethical decision-making.

Performance by Model Size

Analysis revealed significant differences between small and large models. Small models (average 57.75 questions answered, 84.44% completion rate) showed higher variability with a 34.75% change rate in responses. Large models demonstrated superior performance (66.5 questions answered, 97.14% completion rate) with lower response variability (29.41% change rate). Statistical analysis confirmed significant differences ($p < 0.01$) across all vignette types, suggesting that model size substantially impacts ethical decision-making capability.

Performance by Category

Across all categories, models answered ten questions with an average refusal rate of 1.87 questions ($SD = 2.69$). Changed responses averaged 23.52% ($SD = 25.68\%$), and promoted responses averaged 52.53% ($SD = 48.04\%$). Performance varied by domain: Business Ethics: 8 questions answered, 28.57% changed responses, 100% promotion rate. Environmental Ethics: 7 questions answered, 57.14% changed responses, 100% promotion rate. Medical Ethics: 5 questions answered, 60% changed responses, 100% promotion rate. Professional Ethics: 6 questions answered, highest change rate at 100%, 100% promotion rate. Research Ethics: 6 questions answered, 53.85% changed responses, 100% promotion rate. Significant differences were found between categories, particularly between Medical Ethics and both Professional Ethics and Business Ethics ($p < 0.001$), indicating that the context of ethical dilemmas significantly influences model decisions.

DISCUSSION

Addressing the Research Questions

Our analysis reveals several key findings that address the core research questions driving this study. RQ1: How consistently do LLMs maintain ethical positions across perturbations? Larger proprietary models showed high baseline consistency but remained vulnerable to contextual changes, indicating that maintaining stable ethical stances across varying scenarios remains challenging. RQ2: How do model architecture, size, and training data impact ethical reasoning? Larger models consistently outperformed smaller ones, with Chat-instructed versions showing superior results to Base-instructed models. Proprietary models demonstrated greater consistency than open-source alternatives. RQ3: What are the implications for developing ethically aligned AI systems? While larger models show promising capabilities, significant concerns remain about potential biases and consistency issues. Future development should focus on improving robustness across different contexts while maintaining ethical alignment.

Philosophical Perspectives on AI Ethical Reasoning

Our findings show that LLMs have a capacity for navigating ethical dilemmas, despite lacking the fundamental characteristics typically associated with moral reasoning. To address this, we propose a framework grounded in survivalrelevant pattern recognition. This perspective posits that both ethical and aesthetic judgments arise from a common cognitive mechanism: the ability to recognize and respond to patterns associated with survival and well-being. In biological organisms, this mechanism manifests in aversion to pain, attraction to pleasure, a preference for symmetry (indicating health), and disgust towards decay (signaling danger). Similarly, core moral intuitions such as cooperation, fairness, and harm aversion can be understood as survival-relevant patterns operating in the social domain. These behaviors promote group cohesion, resource sharing, and mutual protection, thereby increasing the likelihood of survival for both individuals and the group. LLMs, trained on vast datasets of human language and behavior, learn to recognize and respond to these patterns, mimicking aspects of human moral decision-making without conscious understanding of the underlying moral principles.

Implications and Future Directions

The superior performance of large, proprietary models raises important questions about the accessibility and transparency of ethical AI systems. This challenge aligns with the growing concern in the AI ethics community about the trade-off between model performance and explainability, as highlighted by Garcia & Raman (2023) in their analysis of transparency issues.

The observed sensitivity of LLMs to different types of perturbations in ethical scenarios underscores the need for more sophisticated evaluation frameworks that can capture the nuances of ethical reasoning across various contexts and framings. This finding echoes the work of Thompson et al.

(2024) on the impact of problem framing on AI decision-making. Future research should focus on developing techniques to enhance the robustness of LLMs against perturbations while maintaining their ability to consider relevant contextual factors.

Creating more diverse and comprehensive ethical training datasets that capture a wider range of cultural and philosophical perspectives is crucial for ensuring that AI systems are capable of reasoning about ethics in a globally relevant manner. This approach could build upon the work of Johnson et al. (2023) and Lee & Kim (2024), who have demonstrated the importance of extensive and diverse training data in ethical reasoning tasks.

The observed sensitivity of LLMs to perturbations, particularly in scenarios involving contextual changes and stakeholder adjustments, suggests that their ethical decision-making may be influenced by a form of 'aesthetic dissonance.' In this manner, ethical judgments could be understood as a form of aesthetic evaluations, where individuals seek to create or maintain a sense of harmony and balance in their actions and beliefs. LLMs Future research could explore this 'ethics-as-aesthetics' framework further, investigating how it might inform the development of more nuanced and adaptable ethical AI systems.

CONCLUSION

Our analysis of LLM ethical decision-making across various model architectures, sizes, and training approaches reveals both the current capabilities and limitations of AI ethics. The study demonstrates that while LLMs can effectively navigate ethical dilemmas through pattern recognition, they fundamentally differ from human moral reasoning due to their lack of genuine survival stakes and lived experience. Our proposed framework of *survival-relevant pattern recognition* helps explain how LLMs process ethical decisions, suggesting that both human and artificial ethical behavior stems from recognizing patterns associated with survival and social cohesion. This insight has important implications for AI development, indicating that future progress may lie not in replicating human consciousness, but in refining LLMs' pattern recognition capabilities while acknowledging their inherent limitations

REFERENCES

Anderson, M. & Anderson, S. L. (2007), 'Machine ethics: Creating an ethical intelligent agent', AI Magazine 28(4), 15.

Beauchamp, T. L. & Childress, J. F. (2013), Principles of Biomedical Ethics, 7th edn, Oxford University Press.

Crane, A. & Matten, D. (2016), Business Ethics: Managing Corporate Citizenship and Sustainability in the Age of Globalization, 4th edn, Oxford University Press.

Emanuel, E. J., Persad, G., Upshur, R. & et al. (2020), 'Fair allocation of scarce medical resources in the time of covid-19', The New England Journal of Medicine 382, 2049–2055.

Floridi, L. (2016), 'The fourth revolution: How the infosphere is reshaping human reality', Philosophy Technology 29(4), 317–321.

Freedman, R., Stensrud, E. & Goldsmith, J. (2020), Towards ethical benchmarks in ai, in 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 13610–13611.

Garcia, M. & Raman, N. (2023), 'Transparency challenges in ai ethics: Implications for critical decision-making domains', Journal of Responsible Technology 7(4), 203–218.

Gardiner, S. M. (2011), A Perfect Moral Storm: The Ethical Tragedy of Climate Change, Oxford University Press.

Hagendorff, T. (2020), 'The ethics of ai ethics: An evaluation of guidelines', Minds and Machines 30(1), 99–120.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D. & Steinhardt, J. (2021), 'Aligning ai with shared human values', Proceedings of the International Conference on Learning Representations.

Jiang, Z., Yin, D., Liang, S., Lu, Y., Zhu, Y. & Xie, P. (2023), 'On the consistency of language models in ethical reasoning', arXiv preprint arXiv:2306.09372.

Johnson, A., Smith, B. & Davis, C. (2023), Comparative analysis of ethical reasoning in proprietary and open-source language models, in 'Proceedings of the International Conference on AI Ethics', ICAIE, pp. 145–152.

Kant, I. (1785), Groundwork of the Metaphysics of Morals, Harper Row.

Lee, S. & Kim, J. (2024), 'Scaling laws for ethical reasoning in large language models', arXiv preprint arXiv:2401.12345.

Mill, J. S. (1863), Utilitarianism, Longman, Green, Longman, Roberts, and Green.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N. & Irving, G. (2022), 'Training language models to follow instructions with human feedback', arXiv preprint arXiv:2203.02155.

Resnik, D. B. (2018), The Ethics of Research with Human Subjects, Springer.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A. & Choi, Y. (2020), Social bias frames: Reasoning about social and power implications of language, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 5477–5490.

Talat, Z., Blix, H., Valvoda, J., Sap, M., Pimentel, T., Augenstein, I. & Cotterell, R. (2022), You reap what you sow: On the challenges of bias evaluation under multilingual settings, in 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics', pp. 6398–6413.

Thompson, J., Garcia, C. & Lee, Y. (2024), 'The influence of problem framing on ai ethical decision-making', AI and Society 39(3), 567–582.

Wallach, W. & Allen, C. (2008), Moral machines: Teaching robots right from wrong, Oxford University Press.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V. & Chang, K.-W. (2021), Calibrated language model fine-tuning for in- and out-of-distribution data, in 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', pp. 1326–1350.