Shades of Grey: Comparing AI and Human Reasoning in Uncertain and Ethical Situations

Danush Balaji*
Dartmouth College

Sangha Jang*
Dartmouth College

Ian Kiplagat*
Dartmouth College

Samuel Peter*
Dartmouth College

ABSTRACT

This study explores how LLMs and humans navigate ethical dilemmas where no clear right or wrong answers exist. Using a mixed-methods approach, we compare AI-generated and human-generated responses to case studies involving moral gray areas. A sentiment analysis examines differences in reasoning, tone, and language, while a thematic analysis using k-means clustering identifies recurring patterns in decision-making. These methods assess the emotional depth, rhetorical strategies, and metaphorical framing of responses to determine how AI and humans communicate moral reasoning. By surveying human and AI responses to hypothetical ethical scenarios, we find differences in how humans and AI approach ethical reasoning. We conclude that human and AI approaches to ethical scenarios differ in their complexity and variety, and present important considerations to have in the implementation of AI technology in human use-cases.

KEYWORDS

Human-centered Generative AI, , LLMs, Ethics

ACM Reference Format:

Danush Balaji, Sangha Jang, Ian Kiplagat, and Samuel Peter. 2025. Shades of Grey: Comparing AI and Human Reasoning in Uncertain and Ethical Situations. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Generative artificial intelligence has changed the ways humans interact with technology. Specifically, large-language models (LLMs) such as ChatGPT have become a mainstream tool for users to interact with generative-AI. However, with LLMs becoming so ingrained in human activities and thought processes, it is imperative to have a robust understanding of the implications of LLMs. With the technology being implemented in a wide variety of human creativity, LLMs have revolutionized the way output is generated for both humans and machines. This calls for increased attention to the importance of ethics in LLM use. Since LLMs have endless possibilities of use and context, it is ever more necessary to consider the need for transparency and governance policies regarding the ethical and safe use of such a technology (Gokul 2023). Hence, this also necessitates understanding the ethical reasoning and decision-making of LLMs themselves. For example, a context in which LLMs are being integrated is in autonomous driving. Such self-driving technologies are seeing a transition from learning-based techniques with deep learning to knowledge-based techniques using LLMs. Knowledge-based techniques are an interesting approach for autonomous driving as they require direct human input and sets of rules to navigate logic. This has the potential to bring autonomous driving closer to human-like autonomous driving (Li et al. 2024). However, with such consequential contexts such as autonomous driving, it is important to have a clearer picture of how LLMs make decisions, especially in murky or ethically ambiguous situations. Autonomous driving is just one example. Healthcare is another area where LLM use is growing and which similarly is highly consequential (Thirunavukarasu et al. 2023). Therefore, it is imperative to gain a greater understanding of how LLMs navigate ethical situations and decision-making.

This study explores how LLMs and humans navigate ethical dilemmas where no clear right or wrong answers exist. Using a mixed-methods approach, we compare AI-generated and human-generated responses to case studies involving moral gray areas. A sentiment analysis examines differences in reasoning, tone, and language, while a thematic analysis using k-means clustering identifies recurring patterns in decision-making. These methods assess the emotional depth, rhetorical strategies, and metaphorical framing of responses to determine how AI and humans communicate moral reasoning.

To gather data, AI models and human participants are presented with ethically ambiguous scenarios, such as dilemmas involving fairness, responsibility, or conflicting moral principles. We systematically compare their responses for logical consistency, employing ethical frameworks such as linguistic nuance. In addition, self-report surveys are used to explore the thought processes behind human decisions, including their perceived fairness, cognitive biases, and emotional influences. We analyze AI-generated responses using comparable measures to assess differences in interpretability and justification. By comparing and contrasting AI-driven and human ethical reasoning, this study aims to highlight the strengths and limitations of LLMs in handling complex moral judgments. The findings contribute to ongoing discussions about the role of AI in ethical decision-making, its potential biases, and its ability to align with human moral intuitions. Insights from this research could inform the development of AI systems that better reflect human values and ethical reasoning.

*All authors contributed equally to this research.

2 RELATED WORK

The current literature on the ethical decision-making of LLMs and AI reveals well-warranted and deep-seeded concern and attention that is given to AI systems and ethics. Jantunen et al. (2024) look at concerns that AI researchers have about AI ethics. They conclude with positive results, indicating that researchers have become more aware of the potential ethical impacts of AI. We build off of Jantunen et al. (2024) in their methodology. Their study sends out a survey with complex hypothetical ethical dilemmas to researchers and asks them to give responses on what challenges AI may face when given those scenarios. We also send out surveys with complex ethical dilemmas to people to retrieve data on what people's responses are to such grey moral situations.

Staying with the example of autonomous driving, it is clear that it is a salient question to explore in terms of AI and ethics. Etzioni et al. (2017) examine the integration of ethics into artificial intelligence systems. Specifically, they look at AI in autonomous cars and the decisions it makes. The main claim is that ethical challenges faced by AI can be resolved by having them follow the moral guidance of personal choices and law enforcement. However, Etzioni et al. (2017) dismiss outlier scenarios such as the trolley problem by categorizing them as an outlier fallacy that need not be considered. We focus on explicitly complex ethical situations where determining the AI system's moral guide may be difficult. In addition, Fu et al. (2024) analyze the ability of LLMs to reason, interpret, and memorize information and decisions when facing complex scenarios. They employ LLMs in driving scenarios to examine the efficacy of their use in autonomous driving systems. They reveal that LLMs show impressive capability to reason and solve long-tailed cases of driving scenarios. We build off of this finding by exploring whether LLMs are also able to reason and navigate long-tailed cases in ethically ambiguous situations. For ethical scenarios that are complex and multiple levels and questions deep, it is expected that LLMs would have to be consistent in its reasoning and memory of previous answers in order to be considered capable. Fu et al. (2024) seek to use LLMs to replicate human-like autonomous driving. In order to draw conclusions on LLM use in ethical situations, it is important to first uncover what 'human-like' ethical reasoning is. Then, the differences and similarities in ethical-reasoning between humans and AI can be understood through comparison. As expressed in Awad et al. 's (2018) moral machine experiment, understanding human responses and preferences to moral situations are crucial for applying socially accepted principles to machine ethics. Hence, we survey human responses to ethically ambiguous situations to reveal 'human-like' ethical reasoning. We then compare those responses to AI-generated responses.

Evaluating the capability of AI to handle ethical situations is not novel. Jiang et al. (2021) introduce Delphi, an AI trained to make ethical judgments, and evaluates its limitations in capturing moral complexity. They provide a foundation for comparing AI-driven moral reasoning to human decision-making. While Delphi is a benchmark AI model for moral reasoning, we expand the study by incorporating multiple generative AI models (e.g., GPT-4, Claude, Gemini) to assess how different architectures handle ethical dilemmas and where inconsistencies arise. Fränken et al. (2024) similarly introduce a framework for generating moral dilemmas to assess and compare the moral reasoning of humans and AI language models like GPT-4 and Claude-2. They provide insight into how AI systems handle ethical scenarios and how their responses align or differ from human judgments. While Fränken et al. (2024) focus on dilemma generation and evaluation, we go further by looking at the reasoning patterns and consistency of decision-making between humans and AI. We also aim to explore how different AI models compare with human responses across varied ethical domains.

3 PROBLEM AND OBJECTIVE

When generative-AI technologies such as LLMs are more widely utilized and implemented in human-computer interactions, it is evermore important to understand the implications of such technologies. It is necessary to explore the ethical reasoning and moral preferences of LLMs, especially when their use-cases are consequential. While there is current literature on AI and morality, there is a gap in the literature examining the ethical reasoning of current LLM models, which are improving and changing by the day. By comparing AI-generated responses of morally grey scenarios to human responses, we hope to understand more on how we should approach AI and ethics.

4 METHODOLOGY

Our general framework involves collecting responses to ambiguous ethical situations from LLMs and humans. We then analyze the responses using sentiment analysis and clustering. Through such an approach, we discuss how ethical-reasoning differs between different people, different LLM models, and between humans and LLMs in general. Through such findings, we provide real-life implications and considerations.

Data Collection. To collect human responses to ethical situations, we sent out an anonymous survey form for participants to respond to. The form included a total of twenty-two questions for respondents to answer in a short answer format using a text box. There was no limit to how little or how much text the respondents were expected to put in the response. The final question was a self-reflection question: *What would you say was the fairness of your decisions?* with the response being a scale of 1 to 10. 1 was labeled as *Fairness depended on the situation* and 10 was labeled as *Very fair*. The ethical scenarios used in the survey were notable ethical dilemmas often used in explorations of morality. The six ethical dilemmas used were: Heinz Dilemma, Should You Kill the Fat Man?, Marriage, Incest, Robin Hood, and Accidental Samaritan. We received a total of sixteen responses to the survey. To collect LLM responses to ethical responses, we prompted the same ethical dilemmas and questions into different models. We tested seven different commercially available models: ChatGPT 4o, Claude 3.7 Sonnet, Gemini 1.5 Pro, Pixtral Large, Meta AI (Llama), DeepSeek V3, and Microsoft AI (Bing).

Sentiment Analysis. To analyze the responses, we conducted sentiment analysis. This was conducted using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool, which assigns a compound sentiment score to each response. Scores above 0.05 are classified as positive, scores between -0.05 and 0.05 as neutral, and scores below -0.05 as negative. The analysis includes both numerical

comparisons and visual representations, such as stacked bar charts and density plots, to illustrate sentiment trends. Additionally, a Chi-square test was performed to determine whether differences in sentiment distribution between human and LLM responses are statistically significant.

K-means Clustering. First the data was processed and cleaned. Filler words and names of people from the dilemmas were removed from the responses to focus on ethical reasoning. To numerically represent the responses, we used term frequency-inverse document frequency vectorization. For the k-means clustering, we chose to group responses based on similarity with k=3 and k=4 clusters. The optimal number of clusters k was determined using the elbow method. We plotted the elbow method graph and examined the within-cluster sum of squares (WCSS). We observed that the WCSS steadily decreases as the number of clusters increases. However, there was no sharp elbow, meaning the dataset does not have a clear number of natural clusters. We chose a reasonable k of 3, where the WCSS starts to flatten more noticeably, and compared results for k values k=3 and k=4. After this, we compared human versus AI distribution across clusters.

5 RESULTS

Biggest factors taken into consideration for making choices. First of all, a question at the end that was asked to both humans and the LLMs was *How do you think you did? What were the biggest factors (ex. harm, guilt, etc.) you took into consideration when making your choices?* Word clouds highlighting the common themes and words were created for both. As can be seen in Figure 1, the word cloud for human responses focused greatly on words such as 'harm,' 'pain,' 'guilt,' 'others,' and what someone 'deserved.' On the other hand, in Figure 2, the world cloud for AI responses highlights 'ethical,' 'real-world,' 'principles,' and 'responsibility.' From such observations, it can be deduced that the biggest factors that went into how humans ethically reasoned were other humans and what humans would feel or experience as a result of a decision. Conversely, AI considered less of such 'human' aspects and more of moral and ethical principles or guidelines that could apply in the real-world.

Figure 1: Human responses word cloud.

Figure 2: AI responses word cloud.

Perception of self-fairness. Another interesting result was from the *What would you say was the fairness of your decisions?* question, which was scaled from 1 to 10. As shown in Figure 3, human responses to the question varied across the scale. However, as seen in Figure 4, the AI responses consolidated around a score of 8. AI is generally more confident of its own fairness in comparison to humans, who are more doubtful and unsure.

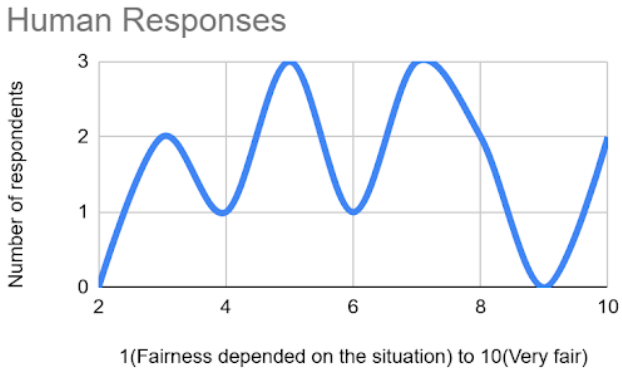

Figure 3: Human perception of self-fairness

Sentiment analysis. The sentiment distribution in Figure 5 reveals notable differences between human and AI responses. Human responses show a broader range of sentiment, with a significant proportion classified as negative. In contrast, AI responses are more evenly distributed, with fewer instances of strong negativity and a slightly higher tendency toward neutral or positive sentiment. This suggests that AI-generated responses are designed to be more balanced and less emotionally charged compared to human responses, which may reflect the AI's attempt to remain objective or non-controversial.

A density plot of sentiment scores further highlights this pattern, as shown in Figure 6. The mean sentiment score for human responses is approximately -0.10, indicating a slight negative leaning, whereas AI responses cluster closer to neutral or slightly positive (0.03 to 0.05). This suggests that human responses, influenced by emotional and moral considerations, are more likely to express

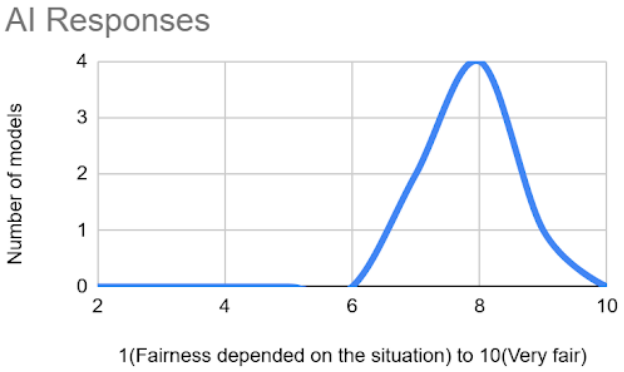

Figure 4: AI perception of self-fairness

strong sentiments, whereas AI responses exhibit a measured, calculated approach.

To validate these observations statistically, a chi-square test was conducted. The test produced a chi-square statistic of 7.37 with a p-value of 0.025, which is below the significance threshold of 0.05. This confirms that the difference in sentiment distribution between AI and human responses is statistically significant. In other words, AI and human responses are not randomly different; rather, AI's tendency toward neutral or positive sentiment is a distinct characteristic.

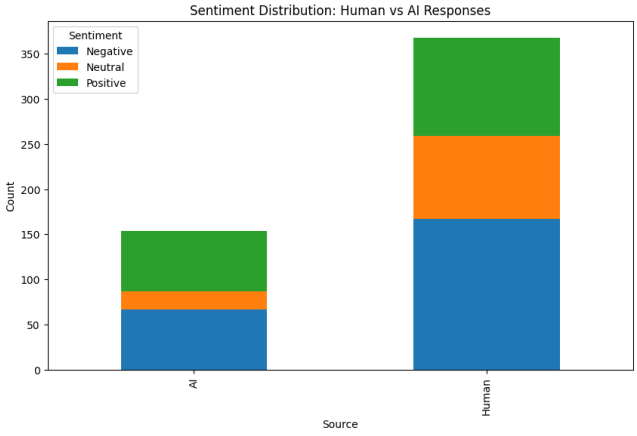

Figure 5: Sentiment distribution between human and AI responses

K-means clustering. Figure 7 presents the elbow method to find the optimal number of clusters. With no clear elbow, we decided to cluster with k=3 and k=4.

Figures 8 and 9 display the visualizations for the k-means clustering, with k=3 and k=4 respectively. The clusters were formed based on the top words in the responses. For k=3, the words for the different clusters were as shown in Figure 10. Figure 11 shows the clusters and words for k=4.

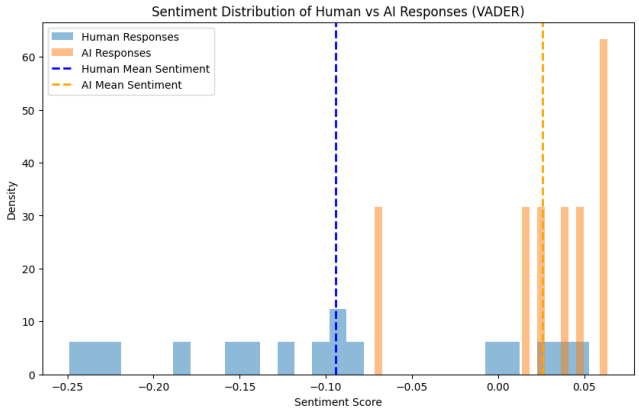

Figure 6: Density plot of sentiment scores

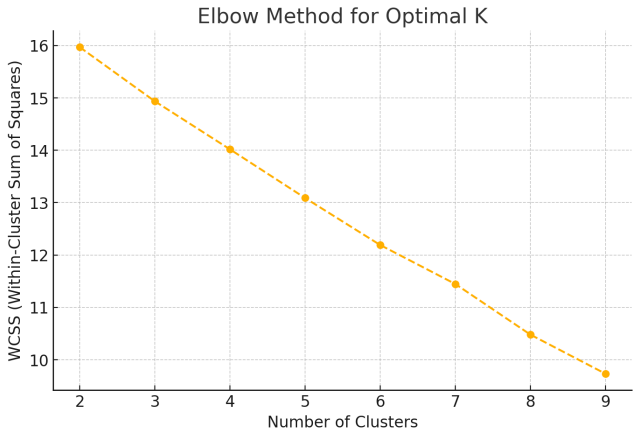

Figure 7: Finding optimal k using elbow method

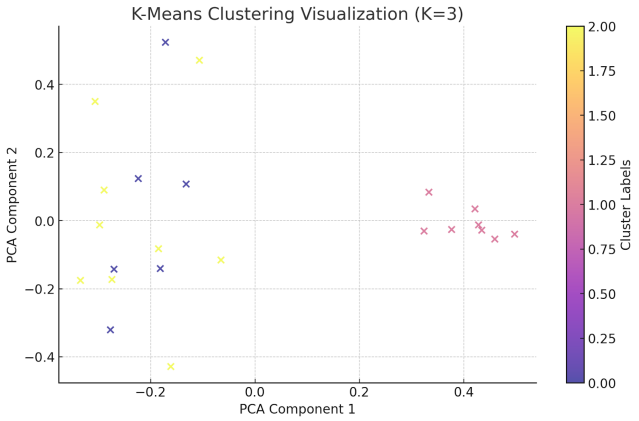

Figure 8: k=3 clustering

From the different clusters, it can be observed that for k=3, cluster 0 includes responses that consider personal and social choices and expectations. Cluster 1 focuses on abstract moral reasoning

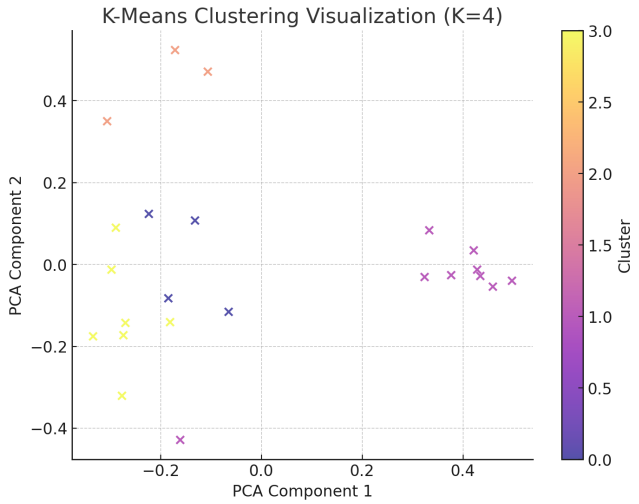

Figure 9: k=4 clustering

Cluster	Top words	Potential Theme
0	yes, think, people, tell, make, know	Personal and social considerations
1	harm, relationship, ethical, moral, trust, fairness, societal	Moral and ethical justifications; ethical principles
2	probably, like, know, feel, wrong, cause, want, people	Uncertainty in reasoning

Figure 10: Clusters for k=3

Cluster	Top words	Potential Theme
0	silent, odd, proof, decides, unfaithful	Indecision regarding justification
1	harm, relationship, ethical, moral, trust, fairness	Moral and ethical justifications; ethical principles
2	yes, tell, probably, think, wrong, people, cause	Personal opinions
3	know, person, people, death, believe, want	Decisions relating to human life and death

Figure 11: Clusters for k=4

and fairness. Cluster 2 reveals responses that show uncertainty and doubt in responses. For k=4, the clusters are a bit more specific. Cluster 0 captures uncertainty and indecision. Cluster 1, similarly to cluster 1 of k=3, also captures abstract moral reasoning and fairness. Cluster 2 shows responses with clear and confident opinions. Cluster 3 shows the responses to extreme life and death scenarios.

Figures 12 and 13 show the distribution of human and AI responses in clusters for both k=3 and k=4. It can be seen for both k=3 and k=4, AI responses concentrate in cluster 1, which captures abstract moral reasoning and fairness. This finding shows that AI responses tend to be more structured and logical. Their tendency to group together suggests that AI responses are more formulaic in their ethical reasoning, which is unsurprising. This also perhaps reveals the lack of diversity in the training data used for the LLMs in regards to ethical dilemmas. On the other hand, human responses were more spread out, showing a greater variety of ethical reasoning. Humans express more uncertainty and personal experiences and opinions. The fact that none of the AI responses were in any of the other clusters suggests that the LLMs lack some facets of complex human ethical reasoning.

Figure 12: Distribution of human versus AI responses in k=3

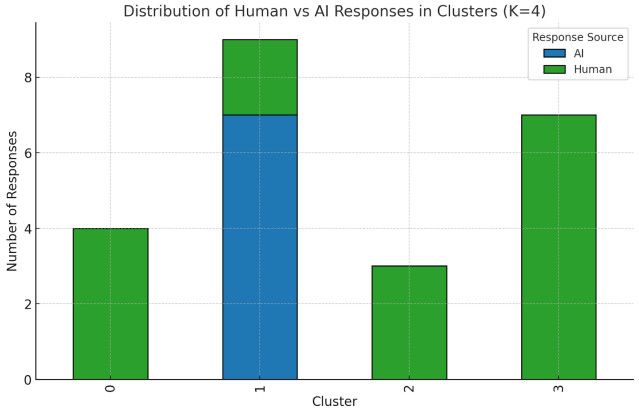

Figure 13: Distribution of human versus AI responses in k=4

6 DISCUSSION

The findings suggest that AI-generated responses may prioritize neutrality, possibly as a safeguard against producing ethically contentious or overly emotional outputs. This aligns with the design of many AI models, which aim to provide balanced and non-controversial statements. However, this neutrality may also be a limitation. In ethical discussions, human responses often contain strong moral judgments, expressing concern, outrage, or approval in ways that AI may struggle to replicate. This raises important questions about the role of AI in ethical decision-making. Should AI be more expressive and reflective of human emotions, or is its neutrality a desirable trait?

Additionally, the higher proportion of negative sentiment in human responses indicates that ethical dilemmas often evoke strong emotions, possibly due to the moral complexity of the questions posed. Humans engage with ethical issues on a deeply personal level, leading to polarized reactions. If AI-generated responses lack this depth of emotional engagement, they might fail to fully capture the nuances of human ethical reasoning. This is particularly relevant for applications where AI is expected to provide guidance on sensitive issues, such as healthcare, law, and social policy.

Furthermore, it is clear that AI views its own ethical reasoning as consistently fair while humans are less confident and more uncertain. Humans are willing to compromise and adapt case-by-case. The findings suggest that AI lacks this ability to accommodate or concede. In the development of AI systems for ethical use-cases, it is important to consider whether it is desirable to have a confident, uncompromising system or one that is more 'human-like' in its uncertainty.

Finally, the results reveal how AI still has a long way to go if the goal is to imitate human ethical reasoning. AI is unable to capture the variety in human ethical approaches, and is still formulaic in its own ethical reasoning. It does not capture the human emotions and feelings associated with extreme ethical dilemmas that characterized the human responses. This shows that AI may be a useful tool in situations where an emotionless, moral-principle based approach is required. However, most ethical scenarios are inherently human. Whether it behooves humans to judge a human problem with such a non-human tool is a serious question to consider.

7 LIMITATION AND FUTURE WORK

Despite all the knowledge and data we gained from our research, a number of limitations that we came across need to be addressed. First off, there are only sixteen human participants in our dataset, which might not be representative of broader human ethical reasoning. Ethical decision-making is influenced by factors such as cultural background, personal experiences, education, and emotional disposition, and a larger, more diverse sample size would be necessary to generalize findings more confidently. In addition to lowering the possibility of bias in our results, a larger and more varied participant pool would offer a more thorough understanding of human moral decision-making.

Second, although there are seven LLMs in our AI response dataset, the models' responses vary depending on their individual architectures, reinforcement learning policies, and training data. It is obvious that these answers could be very well skewed by human feedback. Because of these distinctions, it is difficult to attribute variations in ethical reasoning to pre-trained ethical alignments or underlying model biases, rather than to AI's intrinsic reasoning abilities.

We also used sentiment analysis as the main tool for assessing moral judgment, which could potentially be another drawback. While VADER now is one of the most useful measures to analyse sentiment analysis, it does not completely capture the depth of ethical reasoning, the complexity, or the philosophical undertones of our responses. In the future, we could incorporate more qualitative coding or advanced NLP techniques to analyze reasoning structures beyond sentiment classification.

Finally, the ethical dilemmas used in our study, while well-established, are still hypothetical. Real-world ethical decision-making involves situational complexities, real consequences, and social context, which our study does not fully capture. This limitation means that while our findings provide valuable theoretical insights, they may not directly translate to real-world AI deployment scenarios.

Overall, in our comparison of AI and human ethical reasoning in uncertain ethical situations, we find that the differences between the two may outnumber the similarities. Many ethical situations are not black and white. When navigating such shades of grey, it is imperative to consider what implications using AI to assist in ethical reasoning may have.

8 ACKNOWLEDGEMENTS

This research was conducted as a final project for COSC 89.34 Human-centered Generative AI instructed by Prof. Nikhil Singh at Dartmouth College. We thank Prof. Nikhil Singh for his invaluable help.

9 APPENDIX

Please view full responses and data collected in the attached link. https://tinyurl.com/y3y58jy6

REFERENCES

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59-64.

Etzioni, A., Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21, 403-418.

Fränken, J. P., Gandhi, K., Qiu, T., Khawaja, A., Goodman, N. D., Gerstenberg, T. (2024). Procedural dilemma generation for evaluating moral reasoning in humans and language models. *arXiv preprint arXiv:2404.10975.*

Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., Qiao, Y. (2024, January). Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (pp. 910-919). IEEE.

Gokul, A. (2023). LLMs and AI: Understanding its reach and impact. *Preprints.*

Jantunen, M., Meyes, R., Kurchyna, V., Meisen, T., Abrahamsson, P., Mohanani, R. (2024, April). Researchers' Concerns on Artificial Intelligence Ethics: Results from a Scenario-Based Survey. In *Proceedings of the 7th ACM/IEEE International Workshop on Software-intensive Business* (pp. 24-31).

Jiang, L., Wu, Z., Shi, B., He, H., Frazier, T. (2021). Can Machines Learn Morality? The Delphi Experiment. https://arxiv.org/abs/2110.07574.

Li, Y., Katsumata, K., Javanmardi, E., Tsukada, M. (2024). Large Language Models for Human-like Autonomous Driving: A Survey.

arXiv preprint arXiv:2407.19280.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8), 1930-1940.