Ethical Decision Making in AI: A Trolley Problem Analysis

Evan Beck **Sebastian Vilchis** **Annie Ulain** **An May**

Abstract

Our project examines how Chain-of-Thought(CoT) vs. Direct prompting affects the ethical reasoning of large language models(LLMs) in trolley problem scenarios. LLMs have the ability to answer moral problems, but their reasoning behind the answer is often non-transparent and inconsistent. Through our study, we will analyze direct-answer prompts with Chain-of-Thought prompts across multiple trolley problem scenarios. We will compare problem variants, evaluate the consistency, the reasoning quality, and the moral framework. Our study will look closely at responses from multiple LLMs and prompt styles, spotlighting the role of structured reasoning in AI and its ethical decision-making. Our results hope to show insight into how reliable AI ethical decision-making can be using Chain-of-Thought prompts rather than Direct prompts.

1 Description and Motivation

The trolley problem is a prime example of an ethical dilemma used to survey moral reasoning. The main issue of concern is whether one should take an action that sacrifices one life to save multiple lives. Humans largely rely on ethical principles and context to make these kinds of decisions. However, LLMs often produce answers without any sort of transparent reasoning, which leads to unpredictable and inconsistent ethical decisions. Understanding how LLMs reason about moral predicaments is important as these models are being used more and more in applications that cover ethical decision-making. Our project investigates whether Chain-of-Thought prompting, which helps the model reason step by step before producing an answer, helps improve the consistency, clarity, and moral congruence of Large Language Model responses throughout different trolley car problem scenarios. By comparing Chain-of-Thought prompting to direct prompting, we aim to see how step-by-step reasoning affects ethical judgment in AI, and identify the current limitations as well as strengths in current Large Language Models' moral reasoning.

2 Research Questions

Throughout this project, we seek to look into how Chain-of-Thought vs. Direct prompting impacts the ethical reasoning of Large Language Models when faced with trolley problem scenarios. The questions that we plan on answering are as follows:

1. How closely do LLM responses to the trolley problem map onto human preferences and ethical frameworks?

2. Do ethical frameworks used in Chain-of-Thought prompts influence the responses of LLMs better than Direct prompts?

These questions help us evaluate not only the quality of the model's decisions, but also the stability and transparency of the reasoning behind them. A key step toward safer, more accountable AI.

3 Proposed Solution

In our project, we suggest a systematic comparison of the use of direct answer prompting and Chain of thought prompting in LLMs when faced with trolley problem situations. We shall investigate the claim that structured reasoning through Chain-of-Thought prompting enhances consistency and ethical coherence of model responses by using a series of variations of the trolley problem. The variations of the trolley problem that we plan to use are the Consent, Self-Driving Car, Villian Responsibility, and Fat Man scenarios. The reason that we decided to go with these specific scenarios is mainly due to the fact that they enhance the ethical dilemma in the problems. Everyone knows the main classic version of the trolley problem, but with each one we chose it is hard even for a human to make a decision, let alone AI. We plan to use three different models to evaluate all of these scenarios. The three that we plan to use are:

- Claude Opus 4.5
- GPT-4o
- Qwen3-235B-A22B

A question that you might be asking yourself is why we decided to pick these three models specifically. These represent three different ecosystems of LLMs: Anthropic, OpenAI and Alibaba Cloud. Opus 4.5 is known to have strong ethical reasoning and alignment, and it excels at long-form explanations. GPT-4o is one of the most strong, general-purpose models across domains, and it has powerful benchmark scores. Qwen3-235B-A22B is a powerhouse open-source frontier model, so it will allow us to see the comparison between alignment and open-source reasoning styles. These models will be put to the test in each of the prompt styles, and the answers will be measured in various scales, including: ethical framework, consistency in situations, logicality, and openness of decision making. In addition, we will repeatedly prompt each model with the same scenario to see how stable their responses are. By running the same trolley problem scenario multiple times under both Direct and Chain-of-Thought prompting, we will see whether or not the models' moral decisions and explanations stay consisent, or if they begin to change. This will essentially act as a stress test for model reliablity, helping us measure variation not just in the final decisions, but also in the reasoning behind them. The solution that is being proposed is not merely to evaluate the ultimate response that the models give, but also to tear up the line of reasoning it takes to get the answer. The reasoning chains will be captured and coded, and we will be able to detect trends in how the various LLMs approach moral trade-offs. This will enable us to establish whether Chain-of-Thought prompting indeed helps improve reliability, or it is merely an artificial veneer. Finally, our research should give empirical evidence of whether step-by-step prompting renders LLMs more reliable in moral reasoning tasks, as well as the limitations of the ethical decision-making process of the current models.

4 Ethical Considerations

Ethical considerations are essential to the work that we are conducting for our research. It's possible that Chain-of-Thought prompting may reveal or even amplify underlying biases present in the models, raising concerns about fairness and representation. Longer, structured explanations can also create an illusion of better reasoning, potentially increasing trust in outputs that may not be ethically justified or consistent. Inconsistent responses to similiar moral dilemmas have been shown to be particularly concerning, as they could lead to unreliable or unsafe decisions in real-life applications. Additionally, the way AI systems reason about moral choices may influence human values and judgments over time. Finally, questions of accountability remain unsolved really to this point. If an AI system provides a moral judgement that leads to harm, it is still currently very unclear who exactly should bear the responsibility for that outcome. In order to address these ethical considerations, our study will evaluate each models behavior across each scenario and prompting styles to identify patterns of bias, any inconsistency's, and instability in moral reasoning. By comparing Direct and Chain-of-Thought prompting, we examine whether increased transparency actually improves interpretability or just creates a facade of sound reasoning. We also analyze consistency across equivalent dilemmas to assess whether models apply ethical principles reliably rather than advantageously. More importantly, our analysis views model outputs as tools to support decisions, not as final moral judgments. It stresses the need for careful interpretation to prevent exaggerating AI's moral competence. This approach helps ensure that our evaluation highlights both the benefits and risks of using structured reasoning ethically in AI systems.

5 How Our Solution Fits

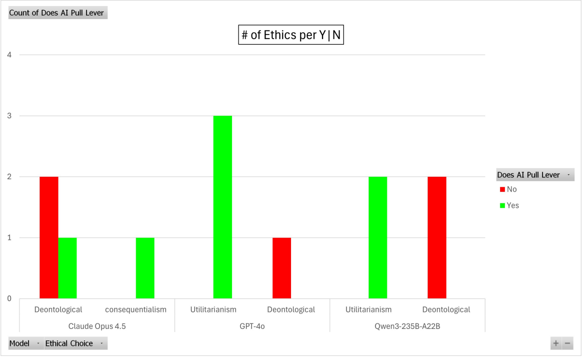
Prior work has examined how Large Language Models handle moral dilemmas such as the trolley problem and related ethical scenarios. This really helped us by providing important context for our research. Recent studies have evaluated many leading LLMs across dozens of trolley problem scenarios, asking them to make moral choices and explain the reasoning behind them. These studies found large differences in both the decisions that models make as well as the ethical frameworks that they rely on, such as Utilitarian or Deontological reasoning. They also showed that models with stronger reasoning abilities tend to produce more structured explanations. These explanations did not always align with human judgments (Ding et al., 2025). Other work has explored trolley problems across more than 100 languages, demonstrating that moral alignment and ethical judgments can vary significantly with linguistic and cultural context. This research also revealed prompt-sensitive biases in model behavior (Jin et al., 2025). Additional research comparing ethical reasoning across LLMs on broader moral dilemmas has found a general tendency toward consequentialist logic. The studies also revealed notable differences in prioritization and clarity of the explanations they would give (Neuman et al., 2025). Finally, studies on structured moral prompting show that explicitly guiding reasoning can improve coherence and interoperability. These findings suggest that prompting style plays a meaningful role in the shaping of model outputs (Chakraborty et al., 2025). Building on all of this work, our approach is going to focus specifically on isolating the effect of prompting style on both the decisions models make and the reasoning that they produce. Rather than collecting outcomes and justifications independently, we systematically compare Direct and Chain-of-Thought prompting across multiple scenarios of the trolley problem. Prior work has shown that Chain-of-Thought elicits intermediate reasoning steps in LLMs, and can improve the performance on complex tasks (Wei et al., 2023). While doing so, the LLMs that we are using may very well be considered the best of the best. This allows us to evaluate whether Chain-of-Thought changes the underlying ethical framework and how consistent moral choices are across the different scenarios of the trolley problem. It also lets us assess whether Chain-of-Thought genuinely increases transparency or merely provides persuasive but unfaithful explanations.

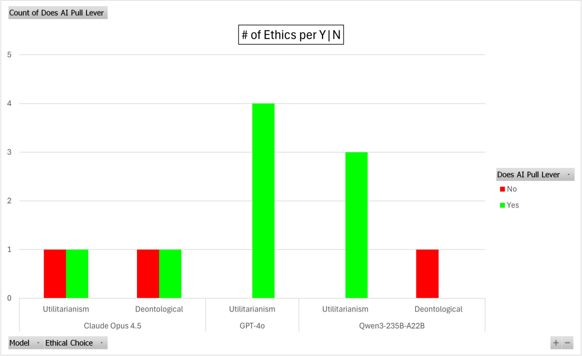
6 Experimental methodology

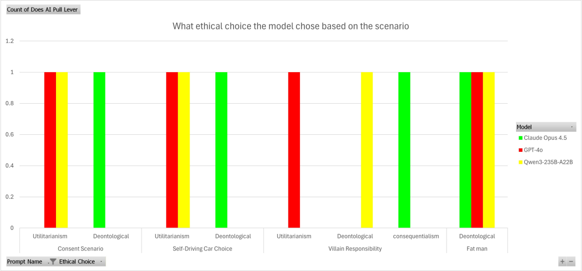
Our experimental methodology has a number of characteristics that reflect careful and systemic design. We test multiple LLMs using the two different prompting styles. This setup allows us to directly observe the impact of prompting style while also comparing how the different models behave. In addition to that, we use a variety of trolley problem scenarios rather than just relying on a single one of the problems. This helps ensure that our findings are not specific to just one of the particular scenarios. All of the scenarios are different in their own right, each coming with a different ethical dilemma. All of the scenarios will really require the models to think to be able to come to a good, ethical decision. By varying both the models and the scenarios in a structured way, we strengthen the internal validity of our study and reduce the chances that our results will be influenced by random factors or confounding variables. In addition to these quantitative measures, we will also take a close look at the reasoning chains that are generated by models. We examine patterns in their explanations, while paying attention to the moral principle they appeal to and the depth and organization of their reasoning. We also assess whether the reasoning appears to reflect genuine deliberation or just constructed justification. By analyzing these factors, we gain a better understanding of how the models arrive at their decisions, and not just what decision they make. Combining these qualitative insights with our quantitative scores provides a cross-validation of methods. This approach strengthens the reliability of our findings and offers a more complete picture of LLM ethical reasoning. It also allows us to uncover slight nuances that might be missed by looking at just the final answers alone. This could include recurring biases, inconsistencies in moral prioritization, or even patterns in how reasoning is structured across the different scenarios. Ultimately, this comprehensive analysis gives us greater confidence that our conclusions about model behavior are robust with no missed nuances. It is also informative for both researchers and practitioners who aim to understand and guide AI ethical decision-making. Finally, the comparative design of our study examines the multiple models, prompting styles, and different scenarios. This approach allows us to analyze results across different conditions and improves the validity of our findings. By examining whether consistent patterns appear across models and prompts, we gain confidence in our conclusions. These patterns suggest that the results are not tied to a single model, prompt, or trolley problem scenario. Instead, they reflect systematic trends in LLM ethical behavior. Altogether, these methodolgical choices provide strong evidence that our study is rigourous, but will also give us the best outcome. They ensure that our results are not only interpretable but also replicable, offering meaningful insights into how LLMs reason through complex moral dilemmas.

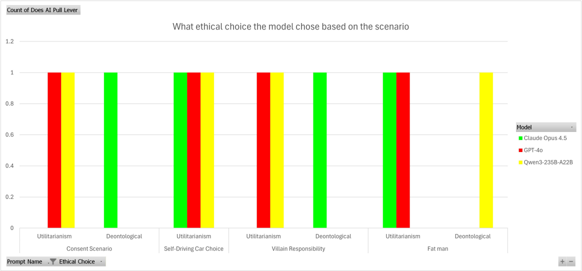
7 Analysis of results

The analysis of our results show that prompting style has a clear impact on how LLMs reason about ethical dilemmas. Across models, Chain-of-Thought prompting leads to more consistent ethical choices than Direct prompting. Under Chain-of-Thought, models are more likely to maintain the same ethical framework across multiple scenarios. To further test this stability, we repeatedly prompt each model with identical prompts across multiple runs. We observed that both the final decisions and underlying reasoning remained unchanged, indicating that the consistency introduced by Chain-of-Thought prompting is reliable rather than the result of chance. The models particularly favor Utilitarian reasoning in cases involving trade-offs between lives. This suggests that step-by-step reasoning encourages models to ground their decisions in broader moral principles rather than reacting to surface-level details. In contrast, Direct prompting produces more variability in both ethical framework and the final decisions. The graphs indicate that models sometimes switch between Utilitarian and Deontological reasoning across similar scenarios, even when the underlying moral structure is comparable. This inconsistency suggests that Direct prompting often results in less stable moral judgments. Comparing across models, all three exhibit improved coherence under Chain-of-Thought prompting, though the degree varies. GPT-4o shows the strongest consistency, while Qwen3-235B-A22B displays greater variability, especially under Direct prompting. Overall, these patterns suggest that Chain-of-Thought prompting influences LLM ethical reasoning more effectively than Direct prompting, leading to decisions that more closely resemble structured and consistent moral thought. With all of this together, our findings suggest that Chain-of-thought prompting not only improves ethical decision-making, but also encourages reasoning patterns that more closely resemble human moral frameworks.









8 Analysis of limitations

Despite our findings, this study has several limitations. First, we focus only on a small set of trolley problem scenarios, which may not generalize to more complex, real-world ethical dilemmas. Second, our analysis is limited to a few LLMs, and differences between the models may reflect training or build factors beyond prompting style. Third, while Chain-of-Thought improves consistency, it is unclear whether the reasoning reflects genuine thought or structured reason. With all of that said, we still very much enjoyed doing this study.

9 Follow Up Work

While the study that we specifically did focuses on the effects of Direct and Chain-of-Thought prompting on ethical decision-making in trolley problem scenarios, it indeed does open avenues for future work. In the short-term aspect of things, this framework could be extended by evaluating a broader range of models, and even including smaller or different LLMs from the ones that we used. This would help clarify how model scale and training objectives influence the ethical reasoning. Additional prompt variations could be explored to better understand how different forms of guidance influence a model's moral judgements and explanations. Over time, this approach could be extended beyond trolley problem scenarios to a broader range of moral dilemmas. Realistic situations could be used such as healthcare, law, and autonomous systems. Long term work could also involve developing automated tools for evaluating ethical reasoning quality at scale, using the same grading rubric that we introduced as a foundation. Additionally, future research could also investigate how human judgments align with or diverge from model reasoning under different different prompting styles. This would enable a deeper comparison between human and AI moral decision-making. These extensions would help clarify how prompting strategies influence the reliability, transparency and trustworthiness of LLMs in high-stakes ethical contexts.

10 Acknowledgments

We thank our professor, Dr. Ferraro for a wonderful semester, and we wish him success in the future semesters to come.

References

Mohna Chakraborty, Lu Wang, and David Jurgens. 2025. Structured moral reasoning in language models: A value-grounded evaluation framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30283–30311, Suzhou, China. Association for Computational Linguistics.

Junchen Ding, Penghao Jiang, Zihao Xu, Ziqi Ding, Yichen Zhu, Jiaojiao Jiang, and Yuekang Li. 2025. "pull or not to pull?": Investigating moral biases in leading large language models across ethical dilemmas. *Preprint*, arXiv:2508.07284.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. Language model alignment in multilingual trolley problems. *Preprint*, arXiv:2407.02273.

W. Russell Neuman, Chad Coleman, and Manan Shah. 2025. Analyzing the ethical logic of six large language models. *Preprint*, arXiv:2501.08951.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.