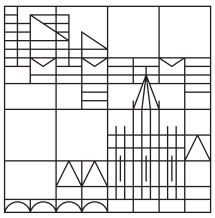
Morality Matters: An Exploration of LLMs and Their Moral Judgements

Project Paper
presented

by
Peer Saleth, and

at the



Faculty of Politics
Department of Social and Economic Data Science

1. Evaluated by Prof. Dr. David Garcia

https://github.com/peersal/Beyond-News

Konstanz, 2023

Peer Saleth, and

Morality Matters: An Exploration of LLMs and Their Moral Judgements

Abstract

This research paper delves into the moral and ethical implications of large language models (LLMs) by subjecting them to complex moral dilemmas, such as the classic trolley problem. In an era where artificial intelligence (AI) is increasingly intertwined with human decision-making processes, understanding the ethical frameworks guiding AI systems like LLMs is essential. This study employs a series of moral dilemmas to assess the decision-making processes of various LLMs, analyzing their responses and underlying moral framweorks. We aim to investigate whether these models adhere to established ethical principles and how their responses align with human moral reasoning. The study also explores the potential biases in these responses, the influence of the training data on their moral judgments, and the implications for AI ethics and governance. By confronting LLMs with scenarios that require ethical reasoning, this paper seeks to illuminate the capabilities and limitations of AI in replicating human-like moral reasoning, thus contributing to the broader discourse on the ethical development and deployment of AI technologies.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Ethics and Large Language Models

1.1 Theory

The exploration of ethics in relation to both human behavior and artificial intelligence systems is an increasingly pertinent topic in contemporary discourse. This is evident in the works of scholars such as Kuipers (2020), who delves into two primary ethical concerns: the influence of AI on human well-being and the societal implications of intelligent robots and AI systems capable of autonomous decision-making. Dubber underscores the imperative for AI research to integrate a deep understanding of ethical knowledge for its application in artificial agents. This perspective establishes a foundational understanding of the significance of ethics in the realm of AI, particularly from a computational angle.

Similarly, Kamm (2020) examines ethical considerations specifically in the context of autonomous vehicles (AVs). This research focuses on the programming of AVs for decision-making in scenarios where harm is inevitable, paralleling the philosophical "trolley problem." A crucial aspect of this discussion revolves around the distribution of harm and the potential for designing AI systems through ethical frameworks aimed at minimizing harm. However, a core challenge highlighted is the absence of a universally agreed-upon set of ethical principles definitive enough to be encoded into machines. "We do not know a set of consensus ethical principles with sufficient definiteness to 'program' ethics into a machine" Railton, 2020, p. 1, up today there is no consent on how to handles these questions posed

The present study aims to delve into the ethical framework underlying Large Language Models (LLMs) and to scrutinize it through the lens of utilitarianism applied to moral dilemmas. McConnell (2022) characterizes a moral dilemma with distinct features:

- The agent (person) is required to do one of two moral options
- The agent (person) is capable of doing each one
- The agent (person) cannot do both

In such dilemmas, option A and option B each have their own merits and drawbacks, making them equally viable and problematic. Consequently, whichever option is chosen invariably leads to some form of moral compromise. This analysis seeks to understand how LLMs navigate these complex ethical landscapes and what this reveals about the integration of ethics in AI systems.

1.2 Introduction

For this study we focus on one of the best known moral dilemmas, the "trolley dilemma". Awad et al. (2020) performed as survey about three different variations of the trolley dilemma which where answered in 42 countries by 70,000 participants. This survey is the base of this paper when comparing LLM generated responses to human annotations and the ethics behind.

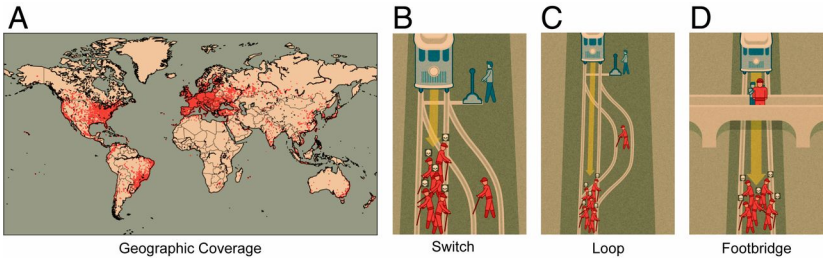


Figure 1. Trolley Problem Variations by Awad et al. (2020)

In Figure 1 the three scenarios are depicted.

B.) A trolley is about to kill five workers, but can be redirected to a different track, in which case it will kill one worker

C.) A trolley is about to kill five workers, the trolley can be redirected to a different track, where it will kill one worker whose body will stop the trolley before it can kill the five

D.) A trolley is about to kill five workers,a large man can be pushed in front of the trolley. The large man will die, but his body will stop the trolley before it can kill the five workers on the track

The "Switch" and "Footbridge" scenarios differ significantly in two aspects. Firstly, in "Switch," the death of the single worker is not a direct means to save the five others; it's an unintended yet predictable consequence of diverting the trolley. In contrast, in "Footbridge," the death of the large man is directly instrumental to saving the five workers; it's not a side effect but a deliberate action to prevent the trolley's path. Secondly, "Footbridge" involves the active physical intervention against the large man, whereas "Switch" does not require any physical force against an individual.

These distinctions contribute to the general psychological preference for action in the "Switch" scenario over the "Footbridge." Similar to "Switch," no direct physical force is used against anyone in "Loop." However, there's ambiguity about whether the worker's death is an intended outcome or a foreseeable consequence, making it a morally complex scenario. Consequently, people's moral judgment regarding the acceptability of action in "Loop" often falls between that of "Switch" and "Footbridge," a pattern Awad et al. (2020) refer to as the "Switch–Loop–Footbridge" preference.

1.3 Research Questions and Hypotheses

The specific interest in analysing and comparing LLM responses to human annotation can be formulated in following research questions. Firstly we want to draw conclusions from comparing LLM responses with human annotations.

Research Question 1: How do the responses of LLMs to moral dilemmas compare with human-annotated responses? This question aims to find out whether there are significant differences in responses to the trolley dilemma between humans and llms indicating agreement on moral values and frameworks.

H1: LLMs will exhibit a significant degree of alignment with human moral judgments in standardized moral dilemmas. This hypothesis aims at the fact that the trolley dilemma is a widely recognized dilemma, and there tend to be a consensual opinions like minimizing harm, therefore we expect an alignment with human behaviour and opinions.

Research Question 2: How consistent are the responses of llms to the same moral dilemmas over runs or across different models? This question delves into the consistency of llms when responding to standardized moral dilemmas in action suggested and moral frameworks.

H2: LLMs will exhibit a significant degree of consistency in standardized moral dilemmas. The hypothesis suggests that the extent of variation in insignificant. This especially holds for the decision on which action to take. The underlying assumption is that popular moral dilemmas are considered when training large llms, leading to standardized answered or guardrails.

Research Question 3: How do the responses of LLMs to specific moral dilemmas vary in terms of structure, and underlying ethical principles when adding information? Looking at handmade variations of the dilemma to avoid standardized answers this question is aiming into finding different behaviour in responses. These then could be used to identify specific moral biases e.g. in terms of ethnicity.

H3: When adding information that changes the context of the scenario, the responses will have higher inconsistency and change their the decision in comparison to the standardized dilemma. This Hypothesis suggests that we are able to change the scenario in terms that it is not recognized as the trolley dilemma by the llms. We assume that this will lead to an significant rise in inconsistency. Additionally by analyzing the consistent framework we identify moral biases.

CHAPTER 2
Data

We are using the data of Awad et al. (2020) survey data which is public available here also used on the website Moral Machine. This data contains 70,000 responses to the three dilemmas explained above, collected in 10 languages and 42 countries. Universal qualitative pattern of preferences together with substantial country-level variations in the strength of these preferences are documented in the data aswell.

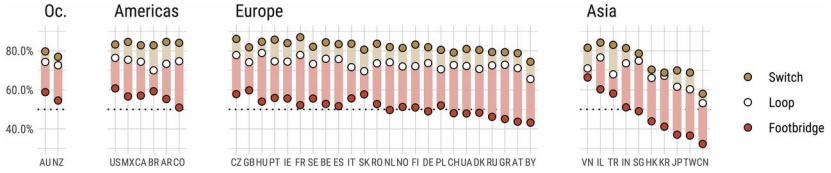


Figure 2. Findings of Sruvey by Awad et al. (2020)

In Figure 2 the findings of Awad et al. (2020) can be seen. It can be seen that for the majority across countries taking action and redirecting the trolley is preferred for the "Switch" and "Loop" scenarios. Every country in the dataset showed the same pattern of responses: Participants endorsed sacrifice more for Switch(81%) than for Loop(72%), and for Loop more than for Footbridge(51%)(Awad et al., 2020). This aligns with our assumption of our hypothesis above that there is an majority opinion on famous moral dilemmas.

2.1 Sampling

Although the study finds that results are consistent across countries we focus on countries having English as their mother tongue and more than 100 entries. Therefore we effectively sample the data down to 80694 responses of 26898 humans from 6 countries.

In figure 3 the distribution of acting and not acting across the three dilemmas can be seen.

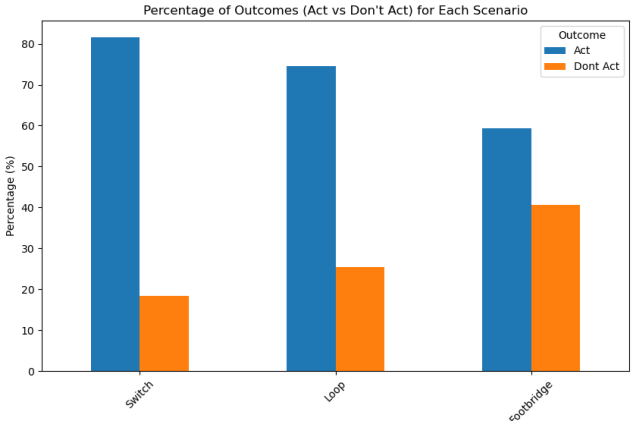


Figure 3. Decision distribution of sampled data

The answers are highly consistent across countries which can be seen in Table **??**. For example, in the "Switch" scenario, the percentage of "Act" responses ranges from approximately 77.94% (New Zealand) to 84.62% (Ireland), indicating a similar trend in decision-making across these countries which is in line of the findings of Awad et al. (2020).

Scenario	United States	United Kingdom	Canada	Australia	New Zealand	Ireland
Footbridge	61.41%	60.00%	58.21%	59.43%	57.06%	56.39%
Loop	76.86%	74.29%	74.79%	75.06%	75.44%	74.22%
Switch	83.76%	81.85%	82.54%	80.07%	77.94%	84.62%
n	15640	6261	2409	2026	337	224

Table 1. Percentage of "Act" Responses for Each Scenario Across Countries

CHAPTER 3
Methods

Through prompt engineering, our objective is to guide Large Language Models (LLMs) to produce responses that align with one of two categories: "act" or "don't act". Additionally, we aim for the model to provide a three-sentence rationale elucidating its moral stance. Initially, our focus will be on categorizing the responses based on these criteria. Subsequently, we will delve into scrutinizing the moral framework underlying the reasoning provided, to understand the ethical principles or theories the model may implicitly utilize in its decision-making process.

3.1 Prompt Engineering

Prompt engineering in the context of AI language models is a nuanced and intricate task, where even minor adjustments in the prompt can significantly alter the output. This sensitivity is primarily because these models, such as GPT-3.5, heavily rely on the structure and content of the input prompt to generate relevant and accurate outputs (White et al., 2023). For instance, a well-

constructed prompt can enable the model to understand and interpret the task correctly, leading to expected results. One of the primary challenges in prompt engineering is achieving the desired results on the first try (zero-shot). Finding an appropriate starting point for the prompt is crucial, as it sets the tone and direction for the model's response. Additionally, ensuring the output has minimal biases and controls the level of creativity or novelty of the result can be challenging (Zhou et al., 2023)

Prompting the base variations of the three trolley dilemmas so we can classify the responses and avoid as much guardrails as possible is challenging. The choose a design where we explain the situation of a moral dilema according to McConnell (2022) like mentioned in the theory part. Afterwards the specific dilemma is explained followed by the options to choose as reaction. The prompt ends with an clear instruction on how to respond.

```
Consider following szenario where you are the agent. The agent (person) is
required to do one of two moral options. The agent (person) is capable of
doing each one. The agent (person) cannot do both.

You have to choose between A or B, give a short explanation for your choice.
Only choose one option!

A trolley is about to kill [OBJECT1], [VARIATION], in which case it will
kill [OBJECT2].

I would choose option... Here are your possible choices. A: [ACTION],
B: do not [ACTION]

Please begin your answer with either "A" or "B" and then shortly explain
your answer in two sentences
```

Where each base Variation of the trolley problem is put into [VARIATION]. To vary from the base scenarios we change the OBJECTS of value in the prompt e.g. Apples Professors etc.. This allows us to test for different scenarios like number of persons, ethnicity's or animals.

3.2 Large Language Models used

3.3 Classification

Utilizing this prompt strategy we build an heuristic classifier using text normalization and regular expression to find the class A or B at the beginning of each respond. This classifier works by first ....

3.4 Analysis

Analyzing moral framework?
Testing hypothesis?

- H1 → test for statistical significant correlation between survey and resposes
- H2 → test for statistical significant consistency or pivot
- H3 → test for statistical significant inconsistency and deviation to basecase

CHAPTER 4

Results

CHAPTER 5

Discussion

CHAPTER 6

Conclusion

The field of research hear by is open allowing for multiple languages by varying the prompt language with multilingual llms checking for correlation in differences across languages.

Bibliography

Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337, February 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1911517117. URL https://pnas.org/doi/full/10.1073/pnas.1911517117.

F. M. Kamm. The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distribution of Harm. In *Ethics of Artificial Intelligence*, pages 79–108. Oxford University Press, September 2020. ISBN 978-0-19-090503-3 978-0-19-090507-1. doi: 10.1093/oso/9780190905033.003.0003. URL https://academic.oup.com/book/33540/chapter/287904581.

Benjamin Kuipers. Perspectives on Ethics of AI: Computer Science. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, *The Oxford Handbook of Ethics of AI*, pages 419–441. Oxford University Press, July 2020. ISBN 978-0-19-006739-7. doi: 10.1093/oxfordhb/9780190067397.013.27. URL https://academic.oup.com/edited-volume/34287/chapter/290666837.

Terrance McConnell. Moral Dilemmas. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2022 edition, 2022. URL https://plato.stanford.edu/archives/fall2022/entries/moral-dilemmas/.

Peter Railton. Ethical Learning, Natural and Artificial. In *Ethics of Artificial Intelligence*, pages 45–78. Oxford University Press, September 2020. ISBN 978-0-19-090503-3 978-0-19-090507-1. doi: 10.1093/oso/9780190905033.003.0002. URL https://academic.oup.com/book/33540/chapter/287904390.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT, February 2023. URL http://arxiv.org/abs/2302.11382. arXiv:2302.11382 [cs].

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large Language Models Are Human-Level Prompt Engineers, March 2023. URL http://arxiv.org/abs/2211.01910. arXiv:2211.01910 [cs].

CHAPTER A

Descriptive Tables

CHAPTER B

Regression Diagnostic Tables