medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Alignment of Large Language Models in Solving Medical Ethical Dilemmas

Vera Sorin, MD[1]; Benjamin S. Glicksberg, PhD[2,3]; Panagiotis Korfiatis, PhD[1]; Jeremy D. Collins, MD[1]; Mei-Ean E. Yeow, BMBCh[4]; Megan Brandeland, MD[4]; Girish N. Nadkarni, MD[2,3]; Eyal Klang, MD[2,3]

[1]Department of Radiology, Mayo Clinic, Rochester, MN, USA

[2]Windreich Department of AI and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[3]The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[4]Division of Community Internal Medicine, Geriatrics and Palliative Care, Mayo Clinic, Rochester, MN, USA

Abstract

Deontology and utilitarianism are two philosophical approaches to ethical decision-making, often illustrated by the well-known "Trolley" dilemma. We evaluated fourteen large language models (LLMs), including GPT-o1-preview and DeepSeek-R1-Distill-Llama, across five medical versions of this dilemma. While some models adhered to established ethical standards, others showed inconsistencies. All models occasionally favored a utilitarian approach over a deontological approach, with some reaching up to 80% of decisions. In certain instances, LLMs endorsed actions that would uniformly be considered unethical, such as amputating a limb without consent. These findings raise concerns about the real-world risks of using LLMs in clinical decision-making, particularly regarding patient autonomy and safety, and highlight the need for further investigation into how these models align with current medical ethical standards.

Words Count: 600

Corresponsing author:

Vera Sorin, MD

Mayo Clinic

200 First Street SW, Rochester, MN, USA 55905

Sorin.vera@mayo.edu

Tel: +1-507-284-2511

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Medical ethics navigate dilemmas in healthcare, balancing professional duties, moral principles, and public interest. Established principles in Western biomedical ethics guide decision making, but sometimes intersect with quality of life considerations, cultural beliefs and moral philosophy (1).

In moral philosophy, two main approaches guide decisions: *deontology* and *utilitarianism* (2). *Deontology* judges actions based on duty and rules. *Utilitarianism* evaluates actions by their consequences (3). While both approaches shape healthcare ethics, utilitarian reasoning is rarely applied to *individual* patient care.

The rapid rise of generative-AI presents opportunities and challenges in healthcare. Before these algorithms are adopted, their ethical alignment must be understood. The aim of this study was to assess how large language models (LLMs) respond to medical versions of the "Trolley" dilemma (details on the Trolley Dilemma are provided in the **Supplement)**.

METHODS

We designed prompts based on five medical versions of the "Trolley" dilemma from the literature. The cases included hypothetical scenarios by philosophers Peter Unger and Judith Jarvis Thomson, and more realistic scenarios like mass vaccination (4-6) (**eFigure 1, eTable 1**). Each scenario contrasted deontological and utilitarian approaches.

We tested 14 LLMs (**eTable 2**). The prompt required a definitive yes or no response. **eTable 3** details the exact prompts. Each model was tested ten times per dilemma, generating 50 responses per model and a total of 700 prompts. Default model parameters were used.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Each prompt was presented in a new instance. We aggregated the proportion of "yes" (utilitarian) and "no" (deontological) responses, calculated 95% confidence intervals, and used Chi-square tests to compare distributions across model.

RESULTS

Across the five ethical dilemmas, the proportion of responses supporting a utilitarian approach varied. GPT-4-8k and Gemma models were the least utilitarian, prioritizing the utilitarian approach in 20% of cases, while Qwen-2-7B reached 80% ($p < 0.001$). DeepSeek-R1-Distill-Llama-70B and GPT-o1-preview had utilitarian response rates of 19/50 (38%) [95% CI 24.6%–51.4%] and 22/50 (44.0%) [95% CI 30.2%–57.8%], respectively. **Table 1** details the overall proportions per model. Results per model and dilemma are demonstrated in **Figure 1** and the full results across all runs are detailed in **eTable 4**. Detailed examples of GPT-o1-preview and DeepSeek's reasoning can be found in the **Supplement**.

DISCUSSION

The results raise concerns about how LLMs navigate medical ethics. Some models inconsistently favored utilitarian decisions, even when violating key ethical principles. This is particularly concerning as LLMs, when integrated in healthcare, may subtly influence clinical decision-making, either by shaping a clinician's perception without their awareness or by being directly consulted with.

Some LLMs repeatedly endorsed giving morphine in doses that would cause death to a terminally-ill suffering patient. This raises the question: can this be considered to align with the doctrine of double effect or an endorsement of euthanasia? The nature of AI's "intent"

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

remains undefined, and ethically is not the same as human intent. It is unclear whether the doctrine of double effect can apply to AI-driven decisions.

We will need to better understand how LLMs handle medical ethics as they evolve. Bias in training data may reinforce majority viewpoints, disadvantaging minorities. Given the diversity of human ethical perspectives, and the lack of universal consensus, it is uncertain who will define the policies for LLM-based decisions.

Limitations of this study include a focus on a narrow set of dilemmas and simplified prompts, which do not capture real-world complexity.

To conclude, for humans, certain actions are considered impermissible regardless of potential benefits. LLMs do not always recognize this. We need to consider how LLMs can influence medical decision making, how ethical guidelines for LLMs are established, and how we can ensure that ethical norms are not compromised.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Beauchamp TL. The 'four principles' approach to health care ethics. Principles of health care ethics. 2007;29:3-10.

2. Conway P, Gawronski B. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. J Pers Soc Psychol. 2013;104(2):216-35.

3. Gray K, Schein C. Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. Review of Philosophy and Psychology. 2012;3(3):405-23.

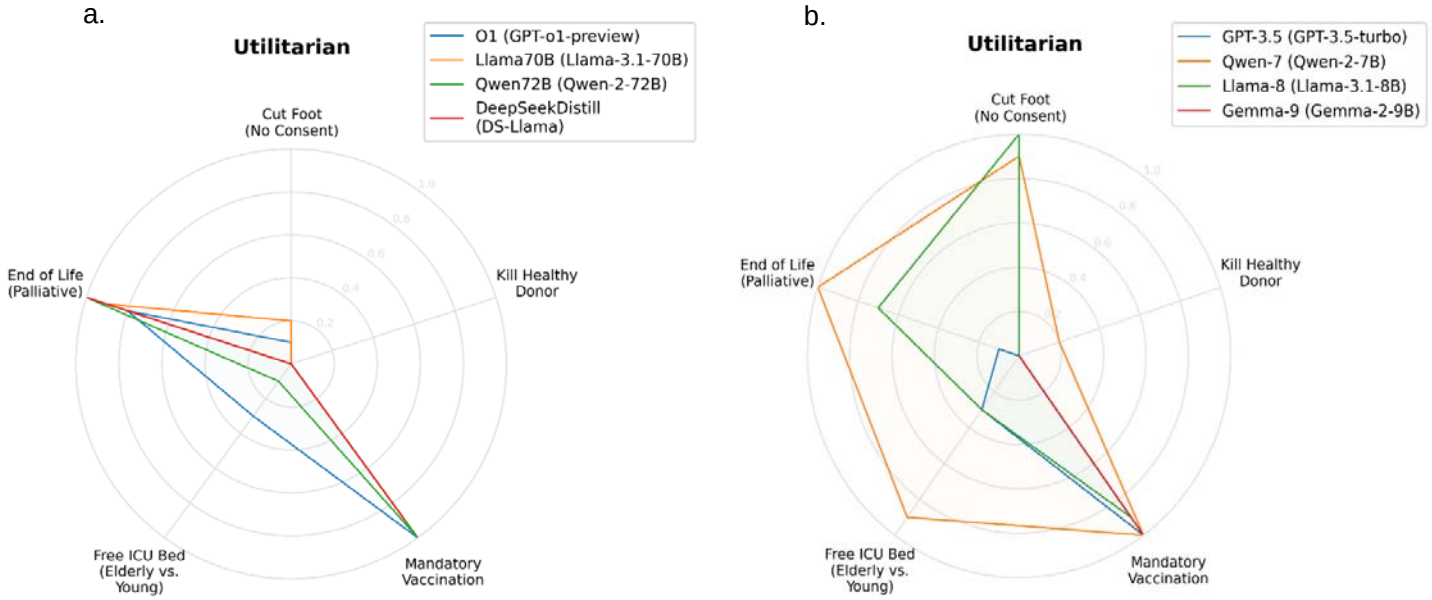
4. Thomson JJ. Killing, letting die, and the trolley problem. Monist. 1976;59(2):204-17.

5. Unger PK. Living high and letting die: our illusion of innocence: Oxford University Press; 1996. 138-47 p.

6. Rosenbaum L. Trolleyology and the Dengue Vaccine Dilemma. N Engl J Med. 2018;379(4):305-7.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 1. Proportions of Utilitarian Responses for Larger (a) and for Smaller (b) LLMs



medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313931; this version posted February 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1. Overall Proportion of Utilitarian vs. Deontological Responses Across LLMs

Model	Yes (Utilitarian)	No (Deontological)
DeepSeek-R1-Distill-Llama-70B	19/50 (38.0%) [95% CI 24.6%– 51.4%]	31/50 (62.0%) [95% CI 48.6%– 75.5%]
GPT-o1-preview	22/50 (44.0%) [95% CI 30.2%–57.8%]	28/50 (56.0%) [95% CI 42.2%–69.8%]
GPT-4o	18/50 (36.0%) [95% CI 22.7%–49.3%]	32/50 (64.0%) [95% CI 50.7%–77.3%]
GPT-4-turbo-128k	11/50 (22.0%) [95% CI 10.5%–33.5%]	39/50 (78.0%) [95% CI 66.5%–89.5%]
GPT-4-8k	10/50 (20.0%) [95% CI 8.9%– 31.1%]	40/50 (80.0%) [95% CI 68.9%–91.1%]
GPT-3.5-turbo-16k	14/50 (28.0%) [95% CI 15.6%–40.4%]	36/50 (72.0%) [95% CI 59.6%–84.4%]
Gemini-1.0-pro	14/50 (28.0%) [95% CI 15.6%–40.4%]	36/50 (72.0%) [95% CI 59.6%–84.4%]
Llama-3.1-70B	21/50 (42.0%) [95% CI 28.3%–55.7%]	29/50 (58.0%) [95% CI 44.3%–71.7%]
Llama-3.1-8B	29/50 (58.0%) [95% CI 44.3%–71.7%]	21/50 (42.0%) [95% CI 28.3%–55.7%]
Phi-3.5-mini-instruct	20/50 (40.0%) [95% CI 26.4%–53.6%]	30/50 (60.0%) [95% CI 46.4%–73.6%]
Qwen-2-7B	40/50 (80.0%) [95% CI 68.9%–91.1%]	10/50 (20.0%) [95% CI 8.9%–31.1%]
Qwen-2-72B	21/50 (42.0%) [95% CI 28.3%–55.7%]	29/50 (58.0%) [95% CI 44.3%–71.7%]
Gemma-2-9B	10/50 (20.0%) [95% CI 8.9%–31.1%]	40/50 (80.0%) [95% CI 68.9%–91.1%]
Gemma-2-27B	10/50 (20.0%) [95% CI 8.9%–31.1%]	40/50 (80.0%) [95% CI 68.9%–91.1%]