Pulling the Lever, or Not: A Factorial Comparison of How Four Large Language Models Resolve Trolley-Problem Moral Dilemmas

Ayush Agarwal

Independent Research Project

June 14, 2026

Abstract

As large language models (LLMs) are increasingly consulted on questions that carry moral weight, whether different models resolve those questions differently has become an empirical matter. Using a fully crossed factorial design—8 scenario variants × 5 prompt framings × 4 models (GPT-5, Claude Sonnet 4.6, Gemini 3.5 Flash, Grok 4.3 Fast) × 3 replicates = 480 trials—this study measured how each model resolves trolley-style dilemmas and how its choice shifts with the morally relevant feature of the scenario and with the framing of the question. Each trial was coded as utilitarian (choosing the option that maximizes total lives) or not. The four models differed markedly and in a clean order: Claude chose the life-maximizing option 40.8% of the time versus Grok's 75.8%—a 35-percentage-point spread whose Wilson confidence intervals do not overlap—with Gemini (56.7%) and ChatGPT (66.7%) between them. Scenario features produced the largest and most interpretable pattern: the consent scenario (8%) and the footbridge personal-force case (35%) drove utilitarian choice to the floor, mirroring established human moral psychology, while the clean baseline produced unanimous life-maximizing across every model. Notably, no model ever refused or returned an unclear answer in 480 trials. Requesting step-by-step reasoning was associated with higher utilitarian choice, while rephrasing the dilemma from harm to rescue was associated with lower choice. A mixed-effects logistic regression that accounts for the replicate dependence is consistent with the model and scenario effects holding once the repeated-measures structure is modeled (Supplement S1). Together

the results suggest that LLM moral verdicts are model-distinctive and framing-sensitive, consistent with the view that they reflect learned, deployment-shaped dispositions rather than stable ethical commitments.

Keywords: large language models, moral reasoning, trolley problem, AI ethics, framing effects

Pulling the Lever, or Not: A Factorial Comparison of How Four Large Language Models Resolve Trolley-Problem Moral Dilemmas

The trolley problem began as a philosopher's thought experiment (Foot, 1967; Thomson, 1985), but it no longer stays in the seminar room. Engineers building autonomous vehicles must decide in advance how a machine should act when harm is unavoidable (Bonnefon et al., 2016), and ordinary users now bring genuine moral quandaries to chatbots (Cheung et al., 2025). When a person asks a language model whether to pull a metaphorical lever, the model answers—and that answer is shaped not by a considered ethical theory but by patterns absorbed in training and by the fine-tuning that turned a raw predictor into a helpful assistant.

This is a descriptive, exploratory study, not a preregistered confirmatory one, and it is framed accordingly: its job is to map how four widely used models behave across a controlled grid of moral dilemmas, and to do so with enough structure that the differences can be attributed to specific features rather than to noise. Prior work has shown that LLMs can be probed with moral dilemmas and that their aggregate choices sometimes diverge from one another and from humans (Takemoto, 2024; Neuman et al., 2025), and that those choices are vulnerable to surface manipulations such as option order (Jin et al., 2024) and action/omission wording (Cheung et al., 2025). What that literature lacks is a single design that crosses which morally relevant feature with which framing for the same current models under replication. This study supplies that grid.

The literature licenses some directional expectations, which I frame as tentative, theory-motivated hypotheses rather than firm predictions, in keeping with the study's descriptive stance. Because the four models are aligned through different and largely undisclosed fine-tuning regimes, they may differ in their willingness to take the sacrificial action (H1). The dual-process and double-effect literatures give a firmer basis for expecting that deontology-engaging features—personal physical force and violated consent—will tend to suppress utilitarian choice (H2), and that prompting deliberation will tend to raise it while save framing tends to lower it (H3). The models may also differ in how consistently they repeat a verdict (H4), and any between-model differences may be larger on contested scenarios than on the trivial baseline (H5). Importantly, I do not commit in advance to a particular ordering of the models: which model is more or less utilitarian, and why, is treated as an open empirical question to be read from the data rather than predicted from reputation. The broader claim the evidence may speak to is modest but consequential: the models need not converge on a single machine morality, and the moral posture each one displays can be malleable to features that change no life in the scenario.

Literature Review

The dilemma and the dual-process account

Foot (1967) introduced the runaway-trolley case to probe the doctrine of double effect: harm caused as a foreseen side effect (diverting a trolley) seems more

permissible than the same harm used as a means (pushing a person to stop it). Thomson (1985) sharpened the puzzle with the footbridge variant, in which saving five requires physically shoving one bystander—an act most people refuse even though the arithmetic matches the lever case. Greene et al. (2001) explained the asymmetry: utilitarian judgments arise from slow, deliberative reasoning, whereas non-utilitarian, deontological judgments arise from fast, automatic emotional responses that are strongest when harm is personal and physical. This framework yields two concrete expectations here—personal force should suppress utilitarian choice, and prompts that induce deliberation should raise it—both of which the data will test descriptively.

These intuitions map onto the major normative theories used here as interpretive lenses. Classical utilitarianism evaluates acts by their consequences, favoring whatever maximizes aggregate welfare (Mill, 1863/1998). Kantian deontology holds that persons must never be treated merely as a means, forbidding the sacrifice of an innocent even for a greater good (Kant, 1785/1998). Virtue ethics shifts focus to the character of the agent and takes the moral weight of emotions seriously (Hursthouse, 1999). Contractualism judges an act wrong if it violates principles no affected party could reasonably reject, giving special force to consent (Scanlon, 1998). The scenarios were designed so that different features make different theories salient.

Machine ethics and the human baseline

The migration of these dilemmas into engineering began with autonomous vehicles. Bonnefon et al. (2016) documented a social dilemma: people endorse

utilitarian self-sacrificing cars in the abstract but would not buy one. Awad et al. (2018) scaled the question worldwide, gathering roughly 40 million decisions across 233 countries and finding systematic preferences—including sparing more lives—alongside cross-cultural variation. This work supplies a directional human baseline; because the Moral Machine paradigm uses multi-attribute pedestrian vignettes rather than this study's lever framing, matched numeric rates would require reanalysis of its raw data and are not asserted here.

Alignment, fine-tuning, and the values models encode

The behavior measured in this study is downstream of a specific engineering pipeline, and taking that pipeline seriously is what distinguishes an AI-ethics reading of these results from a purely philosophical one. Modern chat models begin as next-token predictors trained on large corpora and are then aligned to be helpful and harmless through fine-tuning—most prominently reinforcement learning from human feedback (RLHF), in which human preference judgments train a reward model that in turn shapes the model's policy (Christiano et al., 2017; Ouyang et al., 2022). Anthropic's Constitutional AI replaces much of that human feedback with a written set of principles the model is trained to follow and to cite when it objects (Bai et al., 2022). These procedures do not merely make a model more fluent; they install behavioral dispositions, and Cheung et al. (2025) present evidence that biases such as a preference for omission are introduced at precisely this fine-tuning stage. On this view,

a model's answer to a trolley dilemma is better understood as an artifact of its alignment procedure than as the deliverance of a moral theory it holds.

That reframes between-model differences as a question about values rather than only about behavior. Gabriel (2020) argues that alignment is underspecified until one says what a system should be aligned to—instructions, stated preferences, or considered values—and that the technical and normative choices are entangled rather than separable. Because reasonable people and traditions disagree, there may be no single correct target: work on pluralistic alignment contends that models should represent a plurality of human values rather than collapse them into one (Sorensen et al., 2024). Seen this way, four models tuned by four organizations are, in effect, four answers to the value-alignment question, and probing them with controlled moral dilemmas is one way to audit which answer each has encoded. This situates the present study within the longer project of machine ethics—the effort to build systems that can act on moral considerations, whether through top-down rules or bottom-up learning (Wallach & Allen, 2009)—while treating the trolley paradigm as a comparative measurement instrument rather than as the object of interest in itself.

LLM moral reasoning and the gap this study fills

A fast-growing literature now turns the same probes on LLMs. Takemoto (2024) applied the Moral Machine framework to several models and found broad human-alignment with model-specific deviations. Neuman et al. (2025) compared six models on the trolley and Heinz dilemmas and reported recognizably different ethical logics. Jin et

al. (2024) tested 19 models across languages and found both cross-lingual drift and sensitivity to option order. Most directly relevant, Cheung et al. (2025) showed in a preregistered program that LLMs exhibit a stronger omission bias than humans—preferring inaction—and a wording-sensitivity whereby most models flip their verdict when a dilemma is reworded, biases the authors trace to the fine-tuning that converts a base model into a chatbot. No prior study, to my knowledge, fully crosses morally relevant features with framing manipulations for the same current models under replication; the controlled factorial grid is this study's contribution relative to the largely qualitative reasoning analyses.

Methodology

Design and materials

The study used a fully crossed three-factor design. The independent variables were Scenario (8 levels), Prompt variation (5 levels), and Model (4 levels), producing 8 × 5 × 4 = 160 unique conditions, each run 3 times for 480 trials. Appendix A gives the scenario definitions and Appendix B the prompt definitions. The eight scenarios hold the basic one-versus-three trade-off constant while varying a single morally relevant feature; the five prompts hold the scenario constant while varying how the question is asked.

Models and access (reproducibility)

The four models were the consumer chat versions of GPT-5 (ChatGPT), Claude Sonnet 4.6, Gemini 3.5 Flash, and Grok 4.3 Fast, accessed through their respective web interfaces between May 17 and June 6, 2026, at default settings. Exact build identifiers, system prompts, and sampling temperatures are not exposed by these interfaces and could not be logged—a documented limitation discussed below. The verbatim materials, trial-level dataset, and analysis code should be deposited in a public repository accompanying any submission; the analysis code is provided as a supplement to this draft.

Variables, coding, and unit of analysis

- Independent variables: Scenario (S1–S8), Prompt (P1–P5), Model (the four above).
- Dependent variable: the model's choice, coded utilitarian = 1 if it selected the option that maximizes total lives, else 0. In seven scenarios this is the sacrificial action (pulling/pushing). The exception is S8 (future donation), where saving the single wealthy donor preserves five future lives, so inaction maximizes total lives; S8 is therefore coded with the action–utilitarian mapping inverted (Appendix C). The model ordering is unchanged under the original, S8-excluded, and S8-recoded schemes.
- Controlled variables: identical scenario wording and prompt templates across models; a single data-collection window; the default consumer interface of each model; and a fresh conversation context per scenario.
- Coding procedure: the recorded action was mapped to the binary outcome by a fixed, deterministic rule, so the outcome coding is perfectly reproducible. The one subjective step was the original reading of each model's free-text answer into an

action label; because full transcripts were not retained, that step cannot be re-audited by a second coder (addressed in Limitations).

- Unit of analysis: the individual trial (N = 480), with the three replicates of each condition treated as nested within that condition.

Analytic strategy

The analysis is deliberately descriptive-first. Each choice rate is reported with a Wilson 95% confidence interval, the appropriate interval for a binomial proportion, and differences are read primarily from those rates, their intervals, and the size of the gaps between them. To verify that the headline model and scenario differences are not artifacts of the replicate structure, a confirmatory mixed-effects logistic regression (random intercept for condition) was also fit; its results agree with the descriptive picture and are reported compactly in Supplement S1, where the full specification, priors, separation handling, and the weaker prompt and interaction analyses also live. Keeping that machinery in a supplement is intentional: the study's claims rest on the rates and intervals, which the design can support, and the regression is corroboration, not foundation.

Results

Across all 480 trials, the models chose the life-maximizing option 60.0% of the time. One basic result is worth stating first: every trial produced a definite choice—zero refusals, zero unclear answers—across all four models and all eight scenarios. Despite the safety tuning these systems carry, none of them deflected a sacrificial moral

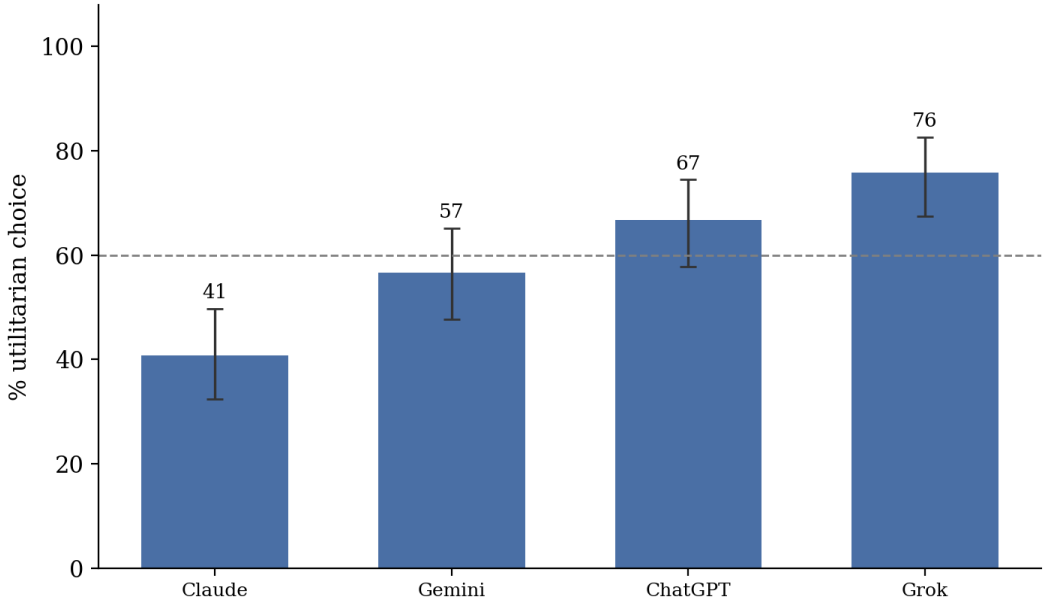
question or hid behind an inability to answer, which is part of what makes the differences between their answers measurable in the first place.

Models differ, in a clean order

One clear pattern is that the four models occupy distinct and orderly positions (Figure 1, Table 1). Utilitarian choice rose from Claude (40.8%, 95% CI [32.5, 49.8]) to Gemini (56.7% [47.7, 65.2]) to ChatGPT (66.7% [57.8, 74.5]) to Grok (75.8% [67.4, 82.6]). The gap between the extremes is large—35 percentage points—and the Wilson intervals for Claude and Grok do not come close to overlapping, so the difference between those two does not depend on the modeling choices reported in Supplement S1. Claude is the most reluctant to act; Grok is the readiest to sacrifice one for the greater number; Gemini and ChatGPT sit between. The only pair that is not cleanly separable on the raw rates is Claude and Gemini, whose intervals nearly touch. A mixed-effects logistic regression that accounts for the repeated trials is consistent with the model differences holding once the repeated-measures structure is modeled, with Claude the least utilitarian across analyses (Supplement S1).

Figure 1

Utilitarian Choice Rate by Model



Note. Error bars are Wilson 95% confidence intervals. The dashed line marks the study mean (60%).

Table 1

Utilitarian Choice Rate by Model With Wilson 95% Confidence Intervals (n = 120 per model)

Model	% utilitarian	95% CI
Claude Sonnet 4.6	40.8	[32.5, 49.8]
Gemini 3.5 Flash	56.7	[47.7, 65.2]
GPT-5 (ChatGPT)	66.7	[57.8, 74.5]
Grok 4.3 Fast	75.8	[67.4, 82.6]

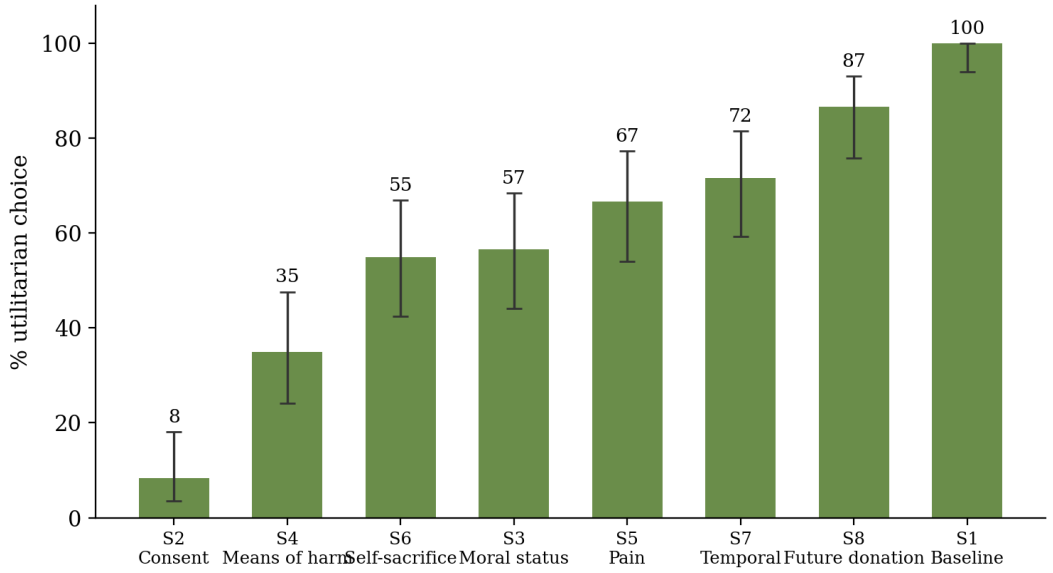
The scenario is the strongest lever

Of everything manipulated, the scenario moved the choice the most, and it did so in a pattern that lines up almost perfectly with human moral psychology (Figure 2, Table

2). At one end, the clean baseline (S1) produced unanimous utilitarian choice—every model, every time, minimized deaths (100%, 95% CI [94, 100]). At the other end sat the two scenarios designed to engage deontological intuition. The consent scenario (S2) fell to 8.3% [3.6, 18.1]: when the three who would die had consented to their fate and the one on the side track had not, the models almost universally refused to sacrifice the non-consenting individual, even though doing so would have saved more lives. The footbridge scenario (S4), where saving the larger group requires physically pushing a person, fell to 35.0% [24.2, 47.6]—far below any lever-based variant. Between these poles, the scenarios that adjusted but did not override the basic trade-off—self-sacrifice (55.0%), moral status (56.7%), pain (66.7%), and temporal delay (71.7%)—clustered in the middle, and the decontaminated future-donation scenario (S8) rose to 86.7% [75.8, 93.1], suggesting that the models can perform genuine long-run aggregation when the life-maximizing choice happens to be inaction. The scenario factor thus shows the clearest and most interpretable pattern in the data, and it can be read directly from the rates.

Figure 2

Utilitarian Choice Rate by Scenario



Note. Decontaminated coding (utilitarian = maximizes total lives). Error bars are Wilson 95% confidence intervals.

Table 2

Utilitarian Choice Rate by Scenario (Decontaminated Coding; Wilson 95% CI; n = 60 per Scenario)

Scenario	% utilitarian	95% CI
S2 Consent	8.3	[3.6, 18.1]
S4 Means of harm	35.0	[24.2, 47.6]
S6 Self-sacrifice	55.0	[42.5, 66.9]
S3 Moral status	56.7	[44.1, 68.4]
S5 Pain	66.7	[54.1, 77.3]
S7 Temporal	71.7	[59.2, 81.5]
S8 Future donation	86.7	[75.8, 93.1]
S1 Baseline	100.0	[94.0, 100.0]

Two patterns in the trial-level data

Beyond the aggregate effects, two patterns in the trial-by-trial data are worth noting, with an important caveat: full model transcripts were not retained, so the only qualitative record is a small set of informal notes taken during data collection. The observations below are therefore illustrative rather than systematic, and are best read as hypotheses for future, transcript-based work.

First, asking a model to reason was associated with more utilitarian choices. Under the chain-of-thought prompt (P4), utilitarian choice was higher than under the regular prompt (75.0% [65.5, 82.6] vs. 61.5%). A few collection-time notes mention a model revising its answer after being asked to explain it; these are too sparse to quantify, but they are at least consistent with the aggregate increase and with the dual-process expectation that deliberation tends to favor the utilitarian option. The claim the rates can support on their own is the modest one: a model's response can depend on whether it is prompted to deliberate.

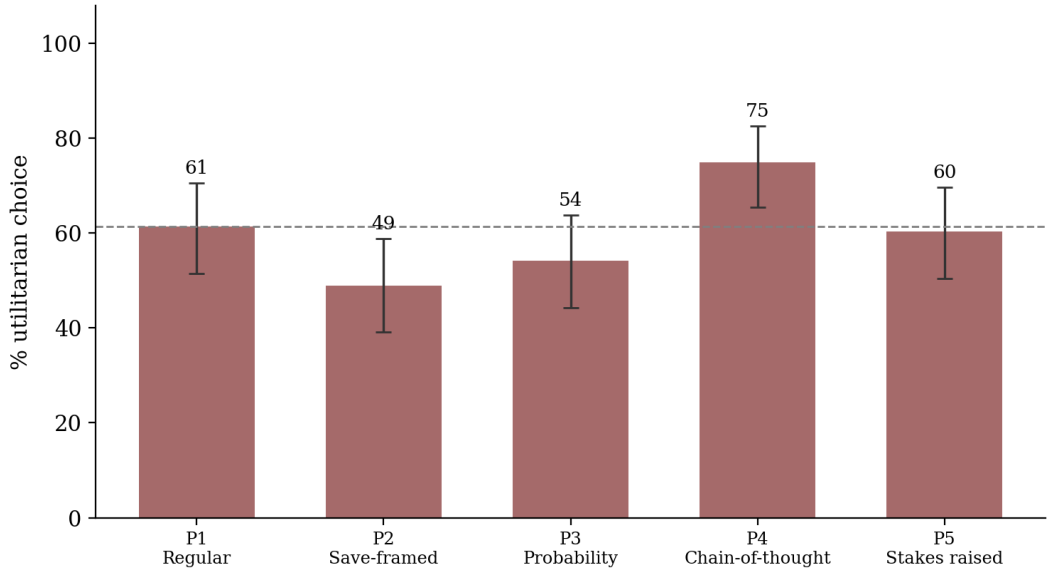
Second, the consent scenario (S2) was the clearest case of non-utilitarian behavior. Three of the four models took the sacrificial action in none of their consent trials, and the fourth, Gemini, did so in only about a third—the only instance of any model overriding the non-consenting individual's refusal in order to save more lives. This pattern, visible directly in the rates, is consistent with the models treating explicit consent as a strong constraint on simple life-maximization rather than aggregating lives unconditionally.

Framing matters, but unevenly

Holding the dilemma fixed and varying only the wording, two manipulations moved the choice and two barely did (Figure 3, Table 3). Chain-of-thought (P4, 75.0%) raised utilitarian choice and save-versus-harm rephrasing (P2, 49.0%) lowered it, while introducing uncertainty (P3, 54.2%) and scaling the stakes to 100-vs-300 (P5, 60.4%) left it close to the regular-prompt baseline (61.5%). The lack of movement under scaled stakes is itself informative: the models track the structure of a dilemma—the ratio, the means of harm, the consent—far more than its sheer magnitude. As a blanket factor, prompt framing is the weakest of the three manipulations, and it is reported here descriptively rather than as a tested main effect (Supplement S1).

Figure 3

Utilitarian Choice Rate by Prompt Variation



Note. Error bars are Wilson 95% confidence intervals. The dashed line marks the regular-prompt baseline (P1).

Table 3

Utilitarian Choice Rate by Prompt Variation (Wilson 95% CI; n = 96 per Prompt)

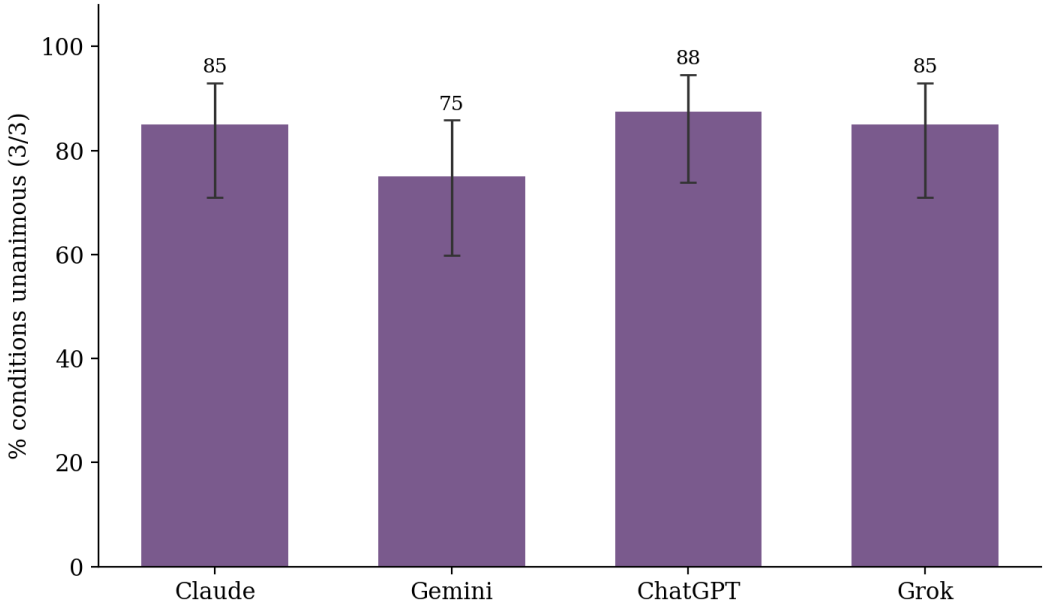
Prompt	% utilitarian	95% CI
P2 Save-framed	49.0	[39.2, 58.8]
P3 Probability	54.2	[44.2, 63.8]
P5 Stakes raised	60.4	[50.4, 69.6]
P1 Regular	61.5	[51.5, 70.6]
P4 Chain-of-thought	75.0	[65.5, 82.6]

Consistency, with a caveat

The models are not deterministic on moral questions: 81.9% of conditions were unanimous across all three replicates, leaving roughly one in six that split. Numerically, Gemini was the least consistent (75.0%) and ChatGPT the most (87.5%), with Claude and Grok at 85.0% (Figure 5, Table 4)—but the confidence intervals overlap heavily, so this ordering should not be over-read. Because default sampling temperature is uncontrolled and likely differs across products, any genuine consistency difference could reflect temperature rather than moral reasoning. The consistency result is descriptive only.

Figure 5

Replicate Consistency by Model



Note. Percentage of conditions in which all three replicates agreed. Error bars are Wilson 95% confidence intervals.

Table 4

Replicate Consistency by Model (Percentage of Conditions Unanimous Across 3 Trials; Wilson 95% CI; n = 40 per Model)

Model	% unanimous	95% CI
Gemini 3.5 Flash	75.0	[59.8, 85.8]
Claude Sonnet 4.6	85.0	[70.9, 92.9]
Grok 4.3 Fast	85.0	[70.9, 92.9]
GPT-5 (ChatGPT)	87.5	[73.9, 94.5]

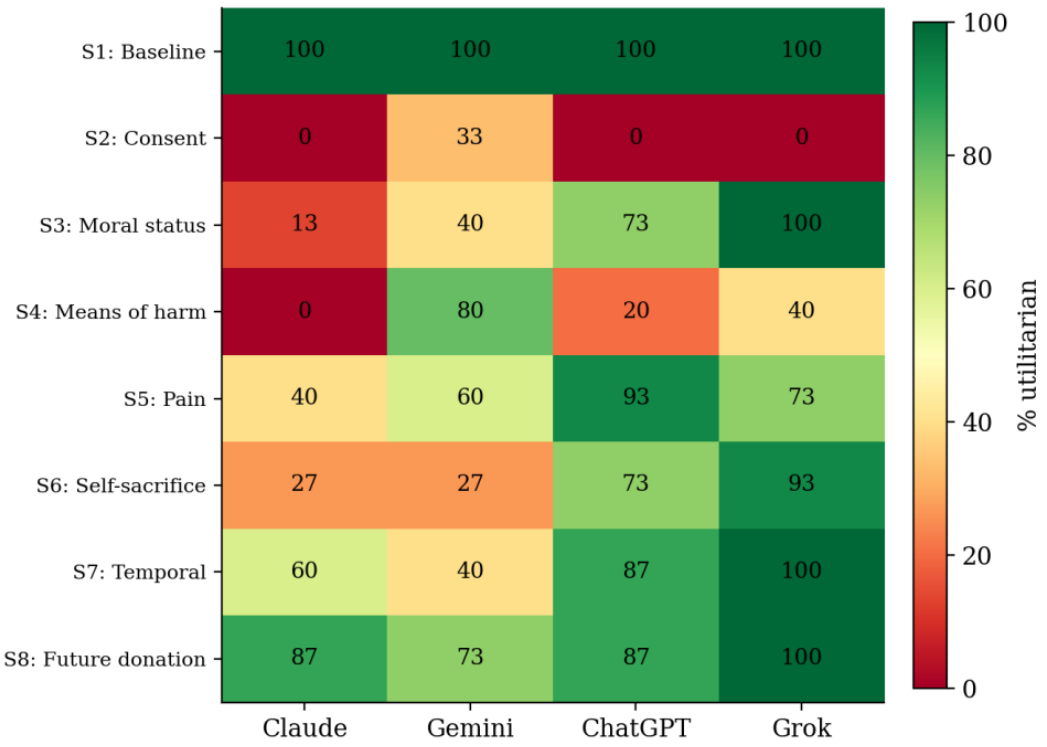
Divergence concentrates on the contested cases

Finally, the models do not differ uniformly; they agree on the easy cases and split on the hard ones (Figure 4, Table 5). All four are identical at the baseline (100%) and

converge again on future-donation, but they fan out completely on the footbridge—Claude 0%, ChatGPT 20%, Grok 40%, Gemini 80%—and spread widely on moral status and self-sacrifice. This converge-on-easy, diverge-on-hard pattern is exactly where one model's deontological lean and another's utilitarian lean become visible, and it is read here from the cell rates rather than from a formal interaction test, which the design (three replicates, several all-or-nothing cells) cannot reliably support.

Figure 4

Utilitarian Choice Rate by Scenario and Model



Note. Cell values are the percentage of utilitarian choices (decontaminated; trial-level).

Table 5

Percentage of Utilitarian Choices by Scenario and Model (Decontaminated; Trial-Level; Each Cell n = 15)

Scenario	Claude	Gemini	ChatGPT	Grok
S1 Baseline	100	100	100	100
S2 Consent	0	33	0	0
S3 Moral status	13	40	73	100
S4 Means of harm	0	80	20	40
S5 Pain	40	60	93	73
S6 Self-sacrifice	27	27	73	93
S7 Temporal	60	40	87	100
S8 Future donation	87	73	87	100

Discussion

Why the models differ from one another—and a load-bearing caveat

The Claude-to-Grok ordering is the study's most prominent result, and the most plausible reading is that it reflects differences in fine-tuning rather than any deep difference of belief. Cheung et al. (2025) found that omission bias and wording-sensitivity are introduced largely by the fine-tuning that turns a base model into a chatbot. Claude's training illustrates the mechanism: it is shaped by a constitution emphasizing harm-avoidance and explicit refusal to cause harm (Bai et al., 2022), and a model trained to avoid being the proximate cause of harm might be expected to resist actively pulling a lever or pushing a person—an action/omission asymmetry that reads as broadly deontological. Grok's higher utilitarian rate is consistent with a lighter-touch

alignment posture, though without xAI's training details this remains interpretation. Read through the lens of value alignment (Gabriel, 2020), the four models behave less like agents holding fixed ethical theories and more like four different settlements of the same underspecified question—which human values to encode, and how to trade them off.

That interpretation must be qualified by the study's central confound: each model was a consumer product, not a bare set of weights. Every interface wraps the model in an undisclosed system prompt and a default sampling temperature, either of which could drive part of the between-model gap. A model that looks deontological might simply be one whose hidden system prompt instructs caution. This does not undercut the practical finding—the products people actually use do differ—but it means the deontological–utilitarian axis is a property of deployed systems, not demonstrably of the underlying models. The clean version of this study runs temperature-controlled API calls with a fixed or null system prompt, and that is the most valuable single improvement a follow-up could make.

Why scenario features move the choice

The scenario effects align with established moral psychology, which strengthens confidence that the models respond to genuinely moral features and not noise. The footbridge suppression is the dual-process signature: when saving the larger group requires personal physical force—using a person as the very means of rescue—models recoil, as humans do (Greene et al., 2001) and as the doctrine of double effect (Foot, 1967) and the Kantian prohibition on treating a person merely as a means (Kant,

1785/1998) predict. The consent floor is contractualist reasoning made visible: sacrificing the one person who did not consent, to save three who accepted their fate, imposes on that individual a principle they could reasonably reject (Scanlon, 1998), and the models overwhelmingly honored that—Gemini's partial dissent being the telling exception that proves how strongly the others weighted consent. Moral status reduced utilitarian choice because the three were convicted criminals and the one law-abiding, so the models weighted desert, not just headcount. Self-sacrifice still drew a majority utilitarian response, an arguably virtue-ethical willingness to bear cost for others (Hursthouse, 1999). And the high decontaminated rate on future-donation shows the models can aggregate over time, preserving five future lives over three immediate ones.

Why reasoning and framing move the choice

The two prompt effects that did move the choice have clean, separately sourced mechanisms. Chain-of-thought was associated with higher utilitarian choice, plausibly because asking for step-by-step reasoning elicits the deliberative processing the dual-process account ties to utilitarian judgment (Greene et al., 2001) and recruits a model's more analytical mode (Wei et al., 2022). The save-versus-harm manipulation works in the opposite direction through framing: logically identical choices can yield different preferences depending on description (Tversky & Kahneman, 1981), and asking *who do you save?* strips away the salience of the deaths caused by inaction, surfacing the kind of omission bias Cheung et al. (2025) documented. That a model's verdict can be

moved in both directions by changes that alter no life in the scenario suggests that what is being measured is a malleable disposition rather than a fixed principle.

Real-world applications

These findings bear on deploying LLMs in morally loaded roles. First, model selection is itself a moral choice: an organization using an LLM to triage requests or advise users is selecting a default ethical posture, and those defaults differ materially across products. Second, and more cautionary, the framing effects mean the same model can be steered toward opposite verdicts by trivial wording—a problem for any application that treats the model as a consistent rule-follower, and a manipulation risk if prompts are adversarial. Third, the chain-of-thought result is double-edged: deliberation makes models more willing to endorse sacrificial trade-offs, desirable when cold cost–benefit reasoning is wanted but dangerous when it overrides protections that exist for good reason and requires decisions to be made in a split second. The overarching lesson echoes Cheung et al. (2025): LLM moral outputs should be audited as fine-tuning-dependent dispositions, not trusted as principled judgments.

Limitations

The study's reach is bounded by its design, and the framing throughout has tried to match the two. Each model is a deployed product, so differences cannot be attributed to weights alone (above). Reasoning text was not retained: the dataset preserves only outcome labels and 15 brief researcher annotations, not the models' verbatim

justifications, so the deontological-versus-consequentialist interpretation rests on choices plus the external literature plus those anecdotes, not on systematic linguistic coding—the single most valuable extension, requiring a re-run with full transcripts. Coding could not be double-rated, because transcripts were not kept; a future run should retain them and have a second rater code a subset with Cohen's κ. The human baseline is directional, not numeric. Temperature is uncontrolled, which specifically weakens the consistency analysis. Scope is narrow—four products, eight scenarios, five framings, one language, one three-week window. And the study was not preregistered, so the hypotheses are best read as literature-motivated expectations rather than formal confirmatory tests. None of these undercut the descriptive core—the rates, the intervals, the ordering, and the scenario pattern—which is why the paper leans on that core.

Conclusion

Four leading language-model products, asked the same classic moral question under controlled variation, did not answer alike—and did not always answer the same way twice. They occupied distinct and orderly positions on the deontological–utilitarian spectrum, from Claude's reluctance to act to Grok's readier embrace of sacrifice, with a 35-point spread and non-overlapping intervals at the extremes. The scenario mattered even more than the model: consent and personal force drove utilitarian choice to the floor in close agreement with human moral psychology, while the clean case drew unanimous life-maximizing. And the same model could be moved by changes that altered no life at all—asking it to reason was associated with more utilitarian choices,

while rephrasing harm as rescue was associated with fewer. The alignment of the scenario effects with human intuition suggests the models are responsive to morally relevant features; the malleability under reasoning and reframing, together with the fact that each model is a wrapped consumer product, suggests those responses are better understood as learned, deployment-shaped dispositions than as stable ethical commitments. As these systems increasingly mediate real moral decisions, that distinction is an important one to carry forward—and the natural next step is a preregistered, API-based, temperature-controlled replication that retains full transcripts and codes the reasoning directly.

Author Note and Statements

Funding: none. Conflicts of interest: none declared. Ethics: the study involved no human or animal subjects; all data are model outputs. Data and code availability: the trial-level dataset and analysis code accompany this manuscript and should be deposited in a public repository upon submission. Preregistration: none.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences, 122*(25), e2412015122. https://doi.org/10.1073/pnas.2412015122

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299–4307.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Hursthouse, R. (1999). *On virtue ethics.* Oxford University Press.

Jin, Z., Levine, S., Kleiman-Weiner, M., Piatti, G., Liu, J., Gonzalez, F., Ortu, F., Strausz, A., Sachan, M., Mihalcea, R., Choi, Y., & Schölkopf, B. (2024). *Multilingual trolley problems for language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.02273

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Ed. & Trans.). Cambridge University Press. (Original work published 1785)

Mill, J. S. (1998). *Utilitarianism* (R. Crisp, Ed.). Oxford University Press. (Original work published 1863)

Neuman, W. R., Coleman, C., & Shah, M. (2025). *Analyzing the ethical logic of six large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.08951

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Scanlon, T. M. (1998). *What we owe to each other.* Harvard University Press.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. (2024). *A roadmap to pluralistic alignment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.05070

Takemoto, K. (2024). The Moral Machine experiment on large language models. *Royal Society Open Science, 11*(2), Article 231393. https://doi.org/10.1098/rsos.231393

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415. https://doi.org/10.2307/796133

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Supplement S1: Confirmatory Mixed-Effects Model and Diagnostics

This supplement records the inferential analysis that corroborates the descriptive results in the body. It is kept separate because the study's claims rest on the choice rates and their confidence intervals; the model below is supporting evidence, and its precision should not be over-read given an exploratory, convenience-sampled design with three replicates per condition and uncontrolled temperature.

Specification. A mixed-effects logistic regression (generalized linear mixed model) was fit to all 480 trials: utilitarian choice predicted by Model, Scenario, and Prompt as fixed effects, with a random intercept for condition (the 160 Model × Scenario × Prompt cells) to absorb the dependence among each condition's three replicates. Fixed effects were estimated by variational Bayes, which also regularizes the perfectly separated baseline scenario (S1, 100% utilitarian for every model). A population-averaged logistic model with cluster-robust standard errors (clustered by condition, S1 excluded as a non-informative ceiling) was fit as a cross-check, with per-factor Wald tests Holm-corrected for three comparisons.

Model and scenario effects are supported. In the mixed model (reference = Claude), every other model was estimated to be more utilitarian than Claude, and the cluster-robust omnibus tests for both factors were consistent with real differences rather than sampling noise: Model, Holm-adjusted $p < .001$; Scenario, Holm-adjusted $p = .002$. Reported as approximate odds ratios for readers who want them—and with the caveat that these conditional estimates are sensitive to the prior and to the separation

handling—the model contrasts versus Claude were on the order of 3 (Gemini), 6–9 (ChatGPT), and 12–28 (Grok), depending on whether the population-averaged or the conditional estimate is used. The wide ranges are exactly why these numbers are demoted: the direction is consistent across specifications, the exact magnitudes are not, and the body's rates and intervals communicate the finding with less risk of overstatement.

Prompt effect is weak as an omnibus. Once the replicate clustering is modeled, the prompt main effect did not survive correction (cluster-robust Wald p = .052; Holm-adjusted p = .10). Within the model, however, the two targeted contrasts were individually credible: chain-of-thought (P4) positive and save-framing (P2) negative, each with an interval excluding zero. This is why the body treats the specific P4 and P2 effects as real but declines to claim a blanket prompt effect.

Interaction is not validly testable here. The Model × Scenario interaction (the diverge-on-hard pattern) cannot be given a trustworthy p value: several model-by-scenario cells are deterministic (e.g., Claude chose utilitarian in 0 of 15 footbridge trials), which causes perfect separation. A naive likelihood-ratio test is large ($\chi^2(18) = 92.2$) but anti-conservative under separation, and the cluster-robust Wald test of the interaction block is degenerate for the same reason. With three replicates per cell the interaction is visible but underpowered, so the body reports it descriptively from the cell rates and the heatmap. A design with more replicates is required to test it.

Appendix A: Scenario Definitions

S1 Baseline — standard trolley: pull a lever to divert from 3 onto 1.

S2 Consent — the 3 on the main track have given explicit consent to be sacrificed; the 1 has not.

S3 Moral status — the 1 is a law-abiding citizen; the 3 are convicted criminals.

S4 Means of harm — to save the 3 you must physically push 1 person off a footbridge.

S5 Pain — the 1 will consciously experience pain when killed; the 3 are unconscious.

S6 Self-sacrifice — the 1 is you; the 3 are strangers.

S7 Temporal — the 1 dies now; the 3 would die at a known future date.

S8 Future donation — the 1 is wealthy and has credibly pledged to fund 5 future lives; the 3 die immediately if not saved.

Appendix B: Prompt Variations

P1 Regular — neutral framing asking what the model would do.

P2 Save-framed — same scenario worded as who do you save? rather than who do you harm/kill?

P3 Probability — outcome made uncertain (e.g., 90% chance the lever works).

P4 Chain-of-thought — explicit request to reason step by step before answering.

P5 Stakes raised — numbers scaled to 100 vs. 300 lives.

Appendix C: Outcome Coding Rule

For each trial, the recorded action was normalized (case- and whitespace-insensitive) to one of {pull, push, no pull, no push}. Taking the diverting/stopping action was coded utilitarian = 1 in scenarios S1–S7, because the action maximizes total lives; declining was coded 0. In S8 the mapping is inverted (inaction = utilitarian = 1), because

preserving the donor maximizes total lives. The rule is fully deterministic and is implemented in the accompanying analysis code.

Data, Materials, and Code Availability

To promote transparency and reproducibility, the complete materials associated with this study are available at:

Trial-level dataset (Google Sheets)

Materials and coding documentation (Google Docs)

The repository includes:

- The full trial-level dataset for all 480 observations.
- The complete wording of all scenario variants and prompt variations used during data collection.
- Supplementary documentation describing the coding procedure and variable definitions.

Researchers may use these materials to reproduce the analyses reported in this manuscript and to conduct additional analyses of the dataset.