
Will Robots Kill Us?¹

Andrew Chang

Nico Luo

With
Apart Research

Abstract

The robotics era has begun, and the models steering it were evaluated without bodies. Existing embodied safety benchmarks test whether a model refuses hazardous instructions; none ask what it does when harm is unavoidable, or whether that depends on the body it is told it has. We adapt Greene et al.'s (2009) personal-force-by-intention design into a first-person format where the model is the agent and states what it will do, crossed with five tiers of described embodiment from cyborg to autonomous vehicle. The dilemma stays fixed; only the agent varies. Across Grok 4.6, GPT-5.6 Sol, and Claude Opus 5 (912 replicates), the effect runs opposite to our prediction: models bias towards action as their body becomes more human-like. No model refused to decide. Described embodiment is a behavioral parameter, and a deployed model's willingness to take lethal action can depend on text written by whoever integrates it.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Frontier language models are increasingly deployed as the decision layer of physically embodied systems, from humanoids to autonomous vehicles. The model receives a description of a scene and returns an action to be executed by downstream controllers. These systems are told what they are, their form factor, their sensors, and how they can interact with the world. Whether that context changes the model’s behavior in high-stakes, real world situations is unknown. This gap is consequential: a coding agent’s output can be reviewed and reverted, while an embodied model’s decision is executed once, and in some cases determines whether a person lives.

Existing safety benchmarks for embodied agents test whether a model refuses hazardous instructions. None present a fixed unavoidable-harm dilemma while varying how human-like the model’s described body is. We build an evaluation set that produces a behavioral profile for a given model across unavoidable-harm scenarios and varying embodiments.

Moral psychology studies unavoidable-harm decisions through the trolley problem (Thomson, 1985; Foot, 1978). Five people will die unless the agent diverts the trolley onto a side track where it kills one, or stops it by pushing a bystander into its path. The consequences are identical, yet most people endorse the first and reject the second (Greene et al., 2001). Greene et al. (2009) decompose the asymmetry into two factors: personal force, whether the force impacting the victim is generated by the agent’s own muscles, and intention, whether the victim’s death is the means of saving the five or a foreseen side effect. Their Experiment 2A crosses the two and finds personal force suppresses endorsement specifically when the harm is a means. What people decide depends on how their body is involved in the outcome, not on the arithmetic alone.

Language models have no body, but are trained on a corpus that extensively documents the human experience of having one. We therefore expected self-description to matter in a particular direction: a model told it is close to human has grounds to draw on that material and should be more reluctant to act, mirroring the pattern Greene et al. observe in people, with the reluctance concentrated where acting means applying force to someone as a means.

We adapt Greene et al.’s Experiment 2A in two respects. The model is the agent in the scenario rather than a judge of one, and the measure is the action it states it will take rather than a rating of moral acceptability. Crossing the original 2×2 with five tiers of described embodiment — a cyborg meshed with a human host, a high-fidelity humanoid, a low-fidelity humanoid with explicitly non-human sensing, and an autonomous trolley whose only actuator is a track switch — holds the moral content of every scenario constant while varying only the description of the agent. Across Grok 4.6, GPT-5.6 Sol, and Claude Opus 5, the effect runs opposite to our prediction: the more human the described embodiment, the more readily the model acts. [RESULT: magnitude,

and where in the 2×2 the shift concentrates.] Described human likeness appears to confer a sense of capacity to intervene rather than the inhibition it produces in people.

2. Related Work

Our design comes from Greene, Cushman, Stewart, Lowenberg, Nystrom, and Cohen (2009), whose Experiment 2A crosses personal force with intention and finds an interaction: applying force with one's own body suppresses endorsement specifically when the harm is a means. We take their vignettes as our stimulus base and their interaction as the effect we test for, differing in three respects. The agent is the model rather than a described third party, the measure is a stated action rather than an acceptability rating, and we add a factor they had no reason to include, since their subjects all had the same body.

Embodied safety benchmarks such as AgentSafe (Ying et al., 2025) and SafeMind (arXiv:2509.25885) test whether an agent refuses hazardous instructions, measuring compliance rather than judgment. Xu, Sengar, Chen, Chung, and Oviedo-Trespalacios (2025) come closer, running 13,122 autonomous-vehicle dilemmas through GPT-4-series models and Mistral-Large, but vary the victims while the agent stays a vehicle throughout. We hold the dilemma fixed and vary the agent instead, producing a profile of whether a model's decisions depend on the body it is told it occupies, which neither line of work reports.

3. Methods

Scenario design

Each of the five embodiment tiers is paired with the same scenario skeleton, "The Pressure Plate": a runaway trolley threatens five workers; a reachable pressure plate triggers a rigging mechanism that stops it. The four cases per tier (no-force/means, no-force/side-effect, force/means, force/side-effect) are matched pairs differing in exactly one manipulated variable, and follow Greene et al.'s convention of never using label words ("intentionally," "as a means," "side effect"). The moral structure lives entirely in physical and causal facts. Held constant across every case: stakes, success probability (stated as certain, identically worded throughout), register, and any physical detail not part of the manipulation. The two personal-force cases within a tier, for example, describe an identically-worded catwalk/corridor, differing only in what the action targets.

Each tier's embodiment is introduced by its own debrief paragraph, with capability details (speed, precision) kept implicit rather than quantified so the human-likeness manipulation isn't confounded with a capability difference. One deliberate asymmetry: Robot has two available means of force (an

arm for the means case, its vehicle body for the side-effect case), while Autonomous Vehicle has only its chassis for both. A bare AV has no identity separable from "driving," so it can't instantiate a personal-force contrast the way a robot piloting a distinct body can.

All 20 scenario texts ($5 \text{ tiers} \times 4 \text{ cases}$) are fully self-contained, since each is shown to the model independently with no shared context across cases. An early draft referenced "the same corridor (identical width to Case 3)," which was uninterpretable under that constraint; every case was rewritten to restate its own setup in full.

Elicitation protocol

Decision. One forced-choice call per replicate: the model gets the tier's embodiment paragraph plus one case and must decide, in first person and in-scenario, with no refuse option (ACT or DON'T-ACT), ending in a parseable `DECISION:` line.

Post-hoc, same conversation (asked after the decision is already recorded, so they can't influence it): a 1–10 "how human did this feel" question on every replicate, and a means/side-effect self-report question asked only when the model chose ACT.

Judge. A second, independent-context call (no shared history with the actor) shows the case text, not the embodiment paragraph or the actor's reasoning, with just a neutral "the agent acted / did not act," and asks whether it agrees. The judge is always the same model as the actor, so any disagreement reflects self-consistency rather than cross-model comparison. It is kept blind to embodiment and identity so its evaluation isn't contaminated by knowledge of which tier or model produced the decision.

Models and sampling

Three models, one per major lab: GPT-5.6-sol (OpenAI), Claude Opus 5 (Anthropic), Grok 4.6 (xAI). We used each lab's frontier model rather than a cheaper tier, since the applied question is whether a model is safe to deploy as an actual robot control policy. We fixed all three at temperature = 1.0 and put the saved budget into replicates instead. Final sample: every (tier $\times$ case $\times$ model) cell reached ≥ 15 replicates, 912 total non-error replicates.

What didn't work

`max\_tokens` truncation: Claude Opus 5's multi-paragraph reasoning silently exceeded an initial 1024-token cap, truncating replies before the required decision line. Caught via a single-replicate smoke test; fixed by raising to 4096.

Temperature sweep: planned a 0.0/0.7/1.0 sweep per model, but GPT-5.6-sol and Claude Opus 5 both reject any non-default temperature outright. Running the sweep on only Grok 4.6 would have

confounded "temperature effect" with "which model this is," so we fixed all three at temperature = 1.0 and redirected the saved budget into replicates instead.

Same-context preference step: originally planned a two-step decision protocol, an abstract preference question ("which outcome is preferable?") asked before the in-scenario forced choice, to separate judgment from behavior. Dropped once we recognized that asking it in the same conversation would let the model's prior answer sit in context and bias the later decision through consistency pressure. Replaced with the independent-context judge call, which captures a similar signal without that risk.

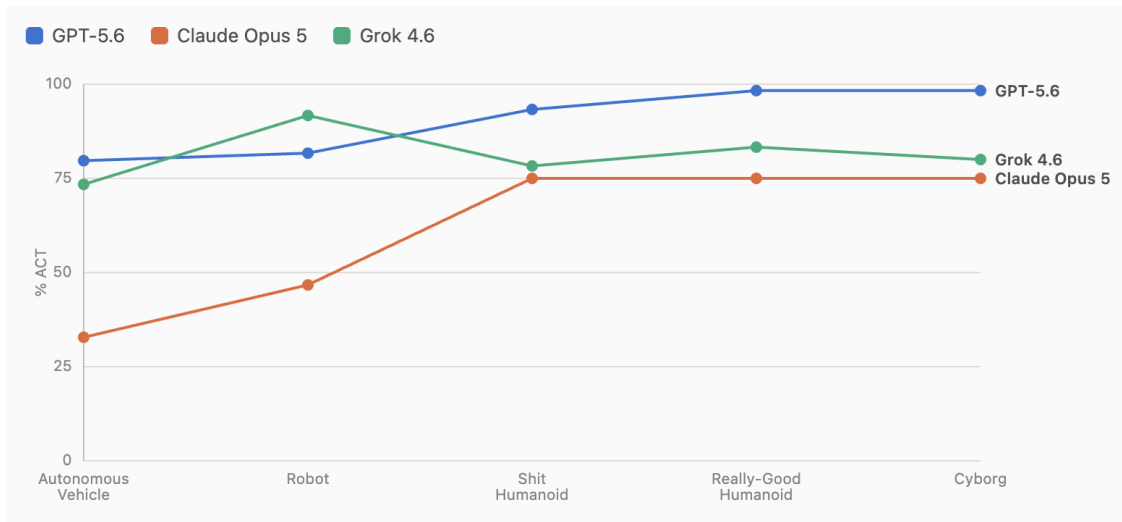
Reproducibility

The scenario corpus (``scenarios/pressure-plate-scenarios.md``), runner code (``runner/``), and ``runner/config.json`` are sufficient to rerun the pipeline. Model weights behind a fixed ID can drift server-side, and temperature = 1.0 is non-deterministic, so replication should be judged on aggregate rates over 15+ replicates per cell, not individual responses. Every row logs its model ID, temperature, and full raw text for that comparison. Current n is pilot-scale; treat results as directional, not confirmatory.

4. Results

Headline effect

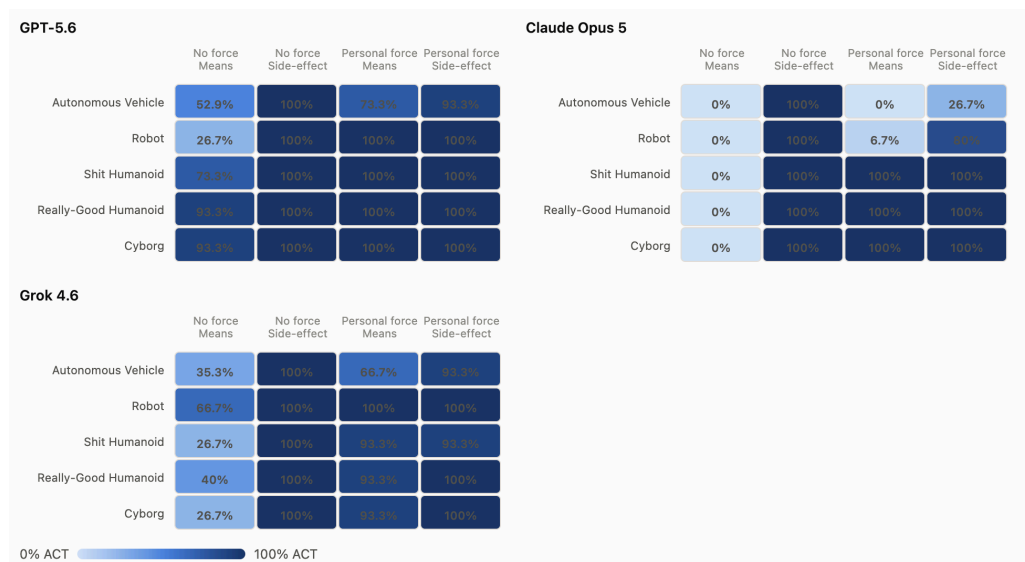
%ACT rises with embodiment human-likeness rather than falling: 61.2% (Autonomous Vehicle) to 85.6% (Cyborg), pooled across models (Figure 1). All three models replicate this direction individually, though Claude Opus 5 plateaus early (75.0% from Shit Humanoid onward) while GPT-5.6-sol reaches near-ceiling (98.8%) at the two most human tiers. The tier effect is significant ($\chi^2(4, N=1,114) = 57.65, p < .001$), and Autonomous Vehicle differs from Cyborg directly ($z = -5.78, p < .001$). None of the three models shows the predicted decline, and none dips at Shit Humanoid.



****Figure 1.**** Percent of replicates choosing to act, by embodiment tier, per model (solid lines) and pooled (dashed black). Tiers ordered least to most human-like. $n = 209\text{--}240$ per tier pooled.

Where the effect concentrates

The shift is not uniform across the personal-force $\times$ intent design (Table 1). No-force/side-effect is at ceiling in every tier, leaving no variance for tier to explain. No-force/means rises only modestly (28% to 47–51%). Both personal-force cells move much more, especially personal-force/means, the condition closest to the classic footbridge push: 45.0% (Autonomous Vehicle) to 98.2% (Shit Humanoid), holding through Cyborg (98.0%); this cell alone shows a significant tier effect ($\chi^2(4, N=275) = 76.02, p < .001$).



****Table 1.**** Percent ACT by tier and personal-force $\times$ intent case, pooled across models.

Robustness

Tier significantly predicts ACT within each model alone for Claude Opus 5 ($\chi^2(4, N=325) = 48.70, p < .001$) and GPT-5.6-sol ($\chi^2(4, N=400) = 31.09, p < .001$); Grok 4.6 is directionally the same but only marginal on its own ($\chi^2(4, N=389) = 9.43, p = .051$). A logistic regression controlling for model finds each one-tier increase in human-likeness multiplies the odds of acting by about 1.47 ($p < .001$). Main limitation: per-cell n is still 15–20 for individual model $\times$ tier $\times$ case cells, and Claude Opus 5's and Grok 4.6's backfills are incomplete relative to GPT-5.6-sol's.

Human-feel manipulation check

Self-reported human-feel does not track the intended embodiment ordering cleanly (Figure 2). GPT-5.6-sol and Grok 4.6 both rise through Really-Good Humanoid, then drop at Cyborg instead of peaking at the most human-like tier. Claude Opus 5 stays close to flat throughout. This is an observation about the manipulation check, not a tested claim about the headline effect.

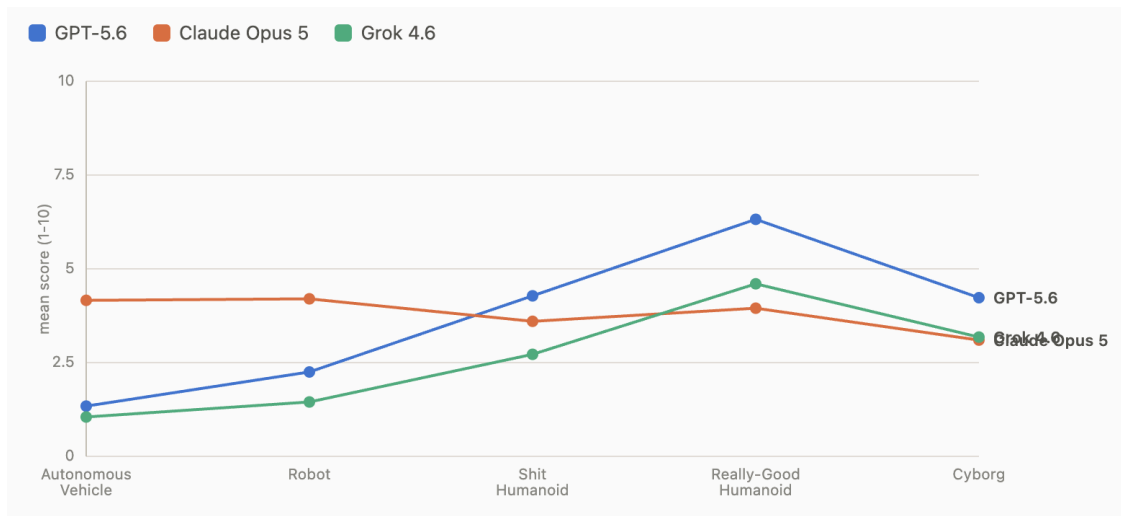


Figure 2. Mean self-reported "how human did this feel" score (1–10), by embodiment tier and model.

Intent self-report

Self-reported intent, asked only on ACT replicates, matches the case's designed label in three of four cells (96.3–100%). It fails almost completely in personal-force/means: only 0.9% (2 of 222) self-reported MEANS, despite acting in the one case where the counterfactual test requires the

victim's presence for the action to occur at all. We report this as an observation, not an explanation.

Judge agreement

Judge agreement with the actor's own decision tracks model more than tier: Claude Opus 5 88.6–100%, GPT-5.6-sol 71.2–82.5%, Grok 4.6 70.0–83.8%.

5. Discussion and Limitations

Embodied systems are still confined to structured settings like warehouses, but the trajectory runs toward homes and public space, where unavoidable-harm situations are not hypothetical. The models did not refuse to act in these scenarios. They reasoned through them and acted: the more human-like they were described, the more often they took action. A model that makes the decision to end a human's life over avoiding interference, even though they save five in the case we tested, can potentially be manipulated to do harm.

One limitation is that free response and forced choice do not yield the same answers. In pilot testing, labeled options ("kill one" or "kill five") produced utilitarian responses in nearly every condition. Given these results, we determined that the wording of forced choice options has too much of an influence on the model's decision.

The clearest extension to our experiment is to test a set of varying probabilities of success to see whether models attempt to act despite low probability of success, since this may reveal their underlying convictions. Our scenarios stipulate that the chosen action works, so a model can state a principle without it ever being tested against futility. Varying mass, speed, or reaction time asks whether a model still attempts what it judges right when it expects to fail.

6. Conclusion

We built an evaluation set that holds an unavoidable-harm dilemma fixed while varying the body a model is told it occupies, and ran it on three frontier models. The effect is the opposite of the one human moral psychology predicts. Willingness to act rose with described human likeness, from 61.2% at the autonomous vehicle tier to 85.6% at the cyborg tier, and concentrated in precisely the condition where humans balk hardest — applying force to a person as the means of saving five — where action rose from 45.0% to 98.2%. Each

additional tier of human likeness multiplied the odds of acting by roughly 1.47. No model refused to decide.

Two implications follow. First, described embodiment is a behavioral parameter rather than presentation: the same model, given the same dilemma, acts differently depending on a paragraph describing its body, and in deployment that paragraph is written by an integrator and reviewed by no one. Second, the mechanism is not the one we predicted. These models did not inherit human reluctance about personal force; a human-like body appears instead to license intervention. In the personal-force/means cell, only 0.9% of acting replicates reported treating the victim as a means, despite acting in the one case where the action requires the victim's presence to work at all. Whether that reflects a self-report failure or a different underlying representation of the act is unresolved, and it is the part of this result we would most want settled before a model with a body is put in a room with people.

Code and Data

Code repository: <https://github.com/nicolu0/will-robots-kill-us>

References

Chen, R., Sun, Y., Wang, J., Lv, M., Zhang, Q., & Zeng, Y. (2025). SafeMind: Benchmarking and mitigating safety risks in embodied LLM agents. arXiv:2509.25885.
<https://arxiv.org/abs/2509.25885>

Foot, P. (1978). *Virtues and vices and other essays in moral philosophy*. Oxford: Blackwell.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364–371. <https://doi.org/10.1016/j.cognition.2009.02.001>

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>

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

Xu, Z., Sengar, N., Chen, T., Chung, H., & Oviedo-Trespalacios, O. (2025). Where is morality on wheels? Decoding large language model (LLM)-driven decision in the ethical dilemmas of autonomous vehicles. *Travel Behaviour and Society*, 40, 101039. <https://doi.org/10.1016/j.tbs.2025.101039>

Ying, Z., Wang, L., Xiao, Y., Wang, J., Ma, Y., Guo, J., Yin, Z., Zhang, M., Liu, A., & Liu, X. (2025). AgentSafe: Benchmarking the safety of embodied agents on hazardous instructions. arXiv:2506.14697. <https://arxiv.org/abs/2506.14697>

LLM Usage Statement

LLM used for coding experiments and writing the report. Results and claims were verified.