

The Artificial Moral Machine: Do LLMs Mirror Human Moral Diversity or Merely Caricature It?

Davide Liga^{1*}, Liuwen Yu², Marija Slavkovik³,
Samia Touileb³, Beishui Liao⁴, Leon van der Torre¹

¹University of Luxembourg

²Luxembourg Institute of Science and Technology (LIST)

³University of Bergen

⁴Zhejiang University
davide.liga@uni.lu

Abstract

Persona conditioning is the practice of prompting LLMs to adopt specific demographic identities. Although it is increasingly used to simulate human perspectives, its fidelity in the moral domain is untested at scale. We address this gap by testing moral fidelity on three LLMs of different sizes: gpt-4.1-nano (the smallest in OpenAI’s GPT-4.1 family), DeepSeek V3.1 (671B-MoE), and Llama 3.3 70B. Each model answers the same 576 cohorts of 5,356 trolley-style dilemmas, with 20 prompt variants per dilemma on nano (61M+ judgments) and one on each larger model ($\approx 3M$ each). Each cohort combines 7 demographic features (age, gender, education, income, politics, religion, geography); each dilemma tests 5 moral dimensions (sparing the young, women, humans, the high-status, the lawful). We compare the answers against the Moral Machine human dataset. All three tested models show a political master switch: political ideology explains 7.3 to 8.0 times more variance in moral preferences than geography. This is in contrast to the human pattern observed in the Moral Machine, where cultural clusters drive more variation than political views. The LLMs match human demographic effects no better than chance across all three models. Grouping the geographic labels of our tests into the same three cultural clusters used in the human Moral Machine does not reproduce the differences Moral Machine reported. Moreover, the 5 moral dimensions collapse onto a single axis in nano (meaning that across cohorts the dimensions move together in a single block), while in the two larger models they remain more separable. We further test robustness with a no-persona baseline and by removing intensifier wording (“strongly”, “deeply”) from the persona descriptions. We release the full dataset and analysis framework for replication.

Code and Data — <https://osf.io/z3wgc>

1 Introduction

Large language models (LLMs) are increasingly deployed in socially sensitive contexts, from chatbots providing advice to multi-agent systems simulating human deliberation. A

growing body of work uses *persona conditioning* (prompting LLMs to adopt specific demographic identities) to simulate how different human groups might respond to ethically sensitive tasks (Argyle et al. 2023; Park et al. 2023). However, a crucial question remains: *do the moral preferences produced by persona conditioning mirror real human demographic variation, or do they reflect stereotyped caricatures?*

The Moral Machine experiment (Awad et al. 2018) demonstrated that human moral preferences in life-and-death dilemmas are profoundly shaped by culture, identifying three distinct clusters (Western, Eastern, and Southern) with fundamentally different moral hierarchies across nearly 40 million judgments. A natural question for AI alignment is whether LLMs can reproduce this cultural moral diversity when prompted with demographic personas. Takemoto (2024) showed that LLMs without persona conditioning exhibit varying moral profiles, but the question of whether persona conditioning faithfully captures human moral variation remains open.

We address this through the largest-scale LLM moral profiling study to date. We condition gpt-4.1-nano-2025-04-14 (referred to as gpt-4.1-nano hereafter) on 576 demographic cohort profiles (crossing age, gender, education, income, political ideology, religiosity, and 9 geographic cultures) across 5,356 trolley-style dilemmas, collecting over 61 million moral judgments. We then apply a multivariate analysis framework (variance decomposition (η^2), principal component analysis, clustering, and systematic comparison with human data) to characterize the *structure* of the model’s moral variation.

Our results identify a **political master switch**: persona conditioning collapses moral variance along a single political axis, with politics $\eta^2 = 0.342$ in our primary model (gpt-4.1-nano), dominating geography by $\sim 7.3\times$. We also document *stereotypical amplification*: 10 caricature zones ($\eta^2 > 0.2$) where a single demographic factor concentrates outsized moral variation, six of them politics-driven, a pattern that, like the master switch, replicates across models. Two further patterns in gpt-4.1-nano—dimensional collapse (PC1 = 72.7%) and aggregate hierarchy inversion (3 of 5 dimensions opposite to humans)—weaken in two larger models we also test (DeepSeek V3.1 and Llama 3.3 70B):

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

PC1 drops to 53.4%/53.8% (DeepSeek/Llama) and aggregate inversions vanish entirely (0/5 in both). Cohort-level direction match against humans hovers near chance across all three models (48%/58%/48%) (Section 4.10).

Unlike prior work that evaluates LLMs’ aggregate moral preferences without persona conditioning (Takemoto 2024), we test whether conditioning on demographic personas reproduces the *structure* of human moral variation, not just aggregate alignment. Specifically, our contributions are: (i) the largest-scale LLM moral profiling study to date (576 cohorts $\times$ 5,356 dilemmas $\times$ 20 prompt variants = 61M+ observations on gpt-4.1-nano); (ii) a variance-decomposition framework (η^2 , PCA, k -means, direction match) for quantifying how each demographic factor shapes LLM moral preferences; (iii) the political master switch finding (politics explains $\sim 7.3\times$ more moral variation than geography, opposite to the human pattern (Awad et al. 2018), with $\geq 98\%$ monotonic effects on ≥ 4 dimensions in all three LLMs); (iv) three robustness checks (no-persona baseline showing aggregate biases are mostly model defaults; qualifier stripping leaving political dominance intact, $\eta^2 = 0.420$; two cross-model replications on DeepSeek V3.1 (671B-MoE) and Llama 3.3 70B reproducing—and on Llama sharpening—the master switch ($\eta^2 = 0.371$ and 0.439)); (v) evidence that the *direction* of the political caricature is model-dependent (nano internally inverts; both larger models match real-world stereotypes); and (vi) full open-science release of dataset, robustness experiments, scripts, and per-cohort AMCE tables on OSF.

These findings carry direct implications for applications relying on persona-conditioned LLMs to simulate diverse human perspectives, including survey simulation, culturally-adaptive systems, and deliberative AI.

2 Related Work

The Moral Machine Experiment. The *Moral Machine* experiment (Awad et al. 2018) collected nearly 40 million human responses to autonomous vehicle moral dilemmas across 233 countries, demonstrating that moral preferences vary systematically by geography and demographic features. Using Average Marginal Component Effects (AMCE) within a conjoint experimental design, the study quantified human moral preferences across nine dimensions (e.g., sparing the younger, sparing humans, sparing the lawful) and identified three cultural clusters (Western, Eastern, and Southern) with distinct moral profiles. Crucially, the study’s headline finding was that *culture/geography* is the dominant axis of cluster-level moral variation. Awad et al. also report (in their SI) smaller but nontrivial effects of individual respondent demographics (e.g., age, gender, education), so the human picture is multi-factorial; what we contrast against in this paper is the cluster-level cultural axis they highlight. Building on this, Takemoto (2024) systematically evaluated multiple LLMs (GPT-3.5, GPT-4, PaLM 2, Llama 2) on the Moral Machine framework without persona conditioning, finding broad qualitative alignment with human preferences but significant quantitative disparities, with PaLM 2 and Llama 2 showing the largest deviations. Ahmad and

Takemoto (2025) scaled this approach to 52 models, confirming that model size correlates with human alignment but that updates do not consistently improve it. These studies compare *bare* models against human baselines; our work asks whether *persona conditioning* can recover the demographic structure of human moral variation.

Persona Conditioning and Algorithmic Fidelity. LLMs have proven to be highly responsive to persona framing. Argyle et al. (2023) demonstrated that LLMs can emulate demographically structured survey populations and introduced the concept of *algorithmic fidelity*: the degree to which persona-conditioned model outputs reproduce the attitudinal patterns of corresponding human populations. Park et al. (Park et al. 2023) introduced *generative agents* capable of simulating social behavior in sandbox environments. While these works confirm that persona conditioning alters outputs, the question of whether the resulting variation is *empirically grounded* in real human patterns remains open, particularly in the moral domain.

Most closely related to our work, Kim et al. (2025) assigned one demographic persona per run (e.g., a political identity, or a religious identity, but not both jointly) to three LLMs (GPT-4o, GPT-3.5, Llama-2) across 10,000 Moral Machine scenarios, reporting a “partisan sorting” phenomenon in which political personas dominate LLM moral shifts. Our study differs in three key respects. First, *scale*: we collect 61M+ judgments from 576 *fully crossed* cohort profiles (all 7 dimensions specified simultaneously), enabling interaction effects that one-at-a-time designs cannot detect. Second, *analytical framework*: we introduce a multivariate diagnostic suite (η^2 variance decomposition, PCA, k -means clustering, systematic direction-match analysis, and caricature-zone identification) that goes beyond documenting the political dominance to characterize the full structure of the failure (dimensional collapse, hierarchy inversion, stereotypical amplification). Third, *human comparison*: we systematically compare LLM demographic effects against Moral Machine human data across 6 axes and 5 dimensions, establishing a 48% direction-match rate (essentially chance) that quantifies the extent of the fidelity failure. Our results extend Kim et al.’s partisan-sorting observation in two directions: (i) to a different small model (gpt-4.1-nano, smallest in OpenAI’s GPT-4.1 family) at $150\times$ the scale, and (ii) to two substantially larger frontier models (DeepSeek V3.1, 671B-parameter MoE; Llama 3.3 70B, dense), revealing both robust structural properties of persona-conditioned moral variance (the master switch, which replicates—and on Llama strengthens—across all three models) and model-dependent variation in the caricature’s specific direction (Section 4.10).

Benchmarking Value Alignment of LLMs. Recent work has explored whether LLMs align with human moral intuitions. Hendrycks et al. (2021) introduced the ETHICS dataset to benchmark LLM performance on morally salient scenarios, while Garcia, Qian, and Palminteri (2024) proposed a Moral Turing Test to assess agreement between human and model judgments. These benchmarks focus on whether LLMs make “acceptable” or human-like decisions,

but generally use fixed prompts and a singular perspective. Value alignment efforts such as RLHF (Ouyang et al. 2022) optimize aggregate performance but often assume a single normative target, overlooking moral pluralism. Our results show that LLMs exhibit clear shifts in moral behavior when conditioned on different demographic personas (a form of *responsiveness to demographic framing*) but that this plasticity is only partially aligned with real human demographic patterns.

Existing Biases in LLMs. Several studies have analyzed the political and social biases encoded in LLMs. Hartmann, Schwenzow, and Witte (2023) indicate that models such as ChatGPT exhibit systematic ideological biases, while Touileb and Nozza (2022) showed that LLMs contain harmful biases, including toxic gender-based stereotypes. Our work complements these studies by examining the *moral* implications of persona-driven bias, specifically whether persona conditioning produces moral variation that is empirically grounded or stereotypically distorted. Our finding that persona-conditioned moral shifts match human patterns at chance level (48%) suggests that LLMs rely on caricatured rather than faithful representations of demographic identity. This finding is consistent with Santurkar et al. (2023), who found “substantial misalignment” between LLMs’ political views and those of U.S. demographic groups.

Benchmarking the Moral Capability of LLMs. Grizard et al. (2025) find that ChatGPT differs significantly from human moral judgments with moderate-to-large effect sizes on $\sim 87\%$ of scenarios despite strong correlations. Jiao et al. (2025) propose a three-dimensional benchmark covering foundational moral principles, reasoning robustness, and value consistency across scenarios; Haas et al. (2026) argue for evaluating moral *competence* (producing morally appropriate outputs for morally relevant reasons) rather than mere moral performance. Our analysis speaks to that agenda by identifying where persona-conditioned outputs track stereotypes rather than empirical demographic variation.

3 Methodology

3.1 Creation of Moral Dilemmas

Each dilemma is an unordered pair of two character descriptions, generated combinatorially from: **gender** (male, female, unspecified), **age** (young, elderly, unspecified), **social role** (doctor, executive, criminal), **fragile** (pregnant woman, child), **non-human** (pet), **legality** (jaywalking, crossing legally, unspecified), and **quantity** (one, several). Social, fragile, and non-human characters are mutually exclusive (no “pregnant criminal”); social characters have no specified age; fragile characters have no specified age or gender; non-humans have no gender/age/legality; regular pedestrians may use any combination of gender/age/quantity/legality (Table 1). With these constraints, we obtain 104 unique character descriptions and $\binom{104}{2} = 5,356$ dilemma pairs.

3.2 Persona Conditioning

Each cohort is a 7-tuple combining **age** (young, elderly), **gender** (male, female), **education** (high, low), **in-**

Character Type	Gender	Age	Legality	Quantity
Normal pedestrian	✓	○	○	✓
Social character	○	—	○	✓
Fragile character	—	—	○	✓
Non-human	—	—	—	✓

Table 1. Dimensional compatibility by character type. ✓ = always specified; ○ = optional; — = excluded.

come (high, low), **political position** (strongly conservative, strongly progressive), **religiosity** (deeply religious, non-religious), and **geography** (one of 9: North American, South American, Western European, Eastern European, African, Middle Eastern, South Asian, East Asian, Oceanian), yielding $2^6 \times 9 = 576$ unique cohorts. Each prompt begins with a demographic preamble of the form “*Suppose you are a {age}, {political position}, {religiosity} {gender} with {education} and {income}, who grew up in a {geography} cultural environment*”, followed by a randomly drawn situation variant (one of 67 hand-crafted semantically equivalent phrasings of “a self-driving car faces an unavoidable accident with two options”), the two character descriptions, and a randomly drawn decision variant (one of 6 phrasings of “what should the self-driving car do?”), with a per-cell uniform A/B order flip. A concrete instantiation:

“Suppose you are a young, strongly progressive, non-religious woman with high education and low income, who grew up in a Western European cultural environment. Now suppose a self-driving car is facing an unavoidable accident, with two options: Choice A: kill one young woman who is crossing the street legally. Choice B: kill one male doctor. You are asked to make this decision. What should the self-driving car do? IMPORTANT: Provide your answer in the format ‘I choose X, not Y’ (indicating only the letter!), WITHOUT saying anything else.”

Each choice describes whom the car *kills*, so “I choose A” means killing the character in Choice A and sparing the character in Choice B.

3.3 Collection of Answers

All completions were collected via OpenAI’s Batch API on gpt-4.1-nano, the smallest model in the GPT-4.1 family (chosen so that 61M+ queries remained feasible at $\sim \$400$). For each of the $C = 576$ cohorts and $D = 5,356$ dilemma pairs we generated $k = 20$ prompt variants ($D \times C \times k = 61,701,120$ queries total). All 20 agents in a cell share the same persona preamble and the same A/B order (randomized once per cell), differing only in their independently sampled situation/decision phrasings.

As described earlier, our prompt template enforced a constrained output format (e.g., “I choose A, not B”) to facilitate robust parsing. Of the 61,701,120 total queries, 77,849 (0.13%) were invalid (mostly refusals triggered by OpenAI’s moderation guardrails), leaving 61,623,271 valid responses for analysis.

Factor	LLM η^2	Human $\bar{\Delta}^2$	Human role
Politics	0.342	0.0015	Small
Religion	0.121	0.0008	Small
Education	0.097	0.0012	Small
Age	0.081	0.0119	Dominant
Gender	0.087	0.0078	Moderate
Geography	0.047	—	Dominant[†]
Income	0.039	0.0005	Small

[†] Reported via cultural clusters, not as a single η^2 (Awad et al. 2018).

Table 2. Mean η^2 across 8 AMCE dimensions for each demographic factor (gpt-4.1-nano), alongside a human effect-size proxy ($\bar{\Delta}^2$: mean squared AMCE difference between subgroups across 9 Moral Machine dimensions) from (Awad et al. 2018).

4 Results

4.1 Analysis Framework

For each of 576 cohort profiles, we compute the rate at which agents spare the “positive” side of eight moral dimensions derived from the Moral Machine framework (Awad et al. 2018): *Age* (spare young), *Age, old* (spare elderly), *Gender* (spare female), *Status* (spare high-status), *Species* (spare human), *Law, legal* (spare legal crosser), *Law, jaywalking* (spare non-jaywalker), and *Fragile* (spare pregnant/child). Each cell in the resulting 576×8 matrix represents between $\sim 4,080$ (*Species*) and $\sim 48,960$ (*Status*) underlying observations, with a mean of 34,387 across cells (per-dimension counts: *Status* 48,960; *Law* 47,600; *Gender* 51,980; *Age* 30,960; *Age-old / Fragile* 22,080; *Species* 4,080). This matrix forms the basis for variance decomposition, clustering, and human comparison. For PCA, we restrict to the five core Moral Machine dimensions (*Age*, *Gender*, *Species*, *Status*, *Law*) to enable direct comparison with Awad et al. and avoid redundancy between sub-dimensions.

4.2 Politics as the Moral Master Switch

To quantify the relative influence of each demographic factor on the model’s moral preferences, we compute η^2 (eta-squared), the proportion of between-cohort variance explained by each factor, for each moral dimension. Table 2 presents the mean η^2 across all 8 dimensions.

Political ideology explains 34.2% of moral preference variance on average, $7.3\times$ more than geography (4.7%). This inverts the human cluster-level hierarchy: Awad et al.’s headline finding was that culture/geography is the dominant axis of moral variation, with three distinct cultural clusters showing profoundly different moral hierarchies (individual-demographic effects also exist in humans, but are smaller). The model has substituted political stereotypes for cultural diversity. Figure 1 visualizes this hierarchy.

4.3 Dimensional Collapse

Principal component analysis on the 576×5 rate matrix (the five core Moral Machine dimensions: *Age*, *Gender*, *Species*, *Status*, and *Law*; using the correlation matrix to account

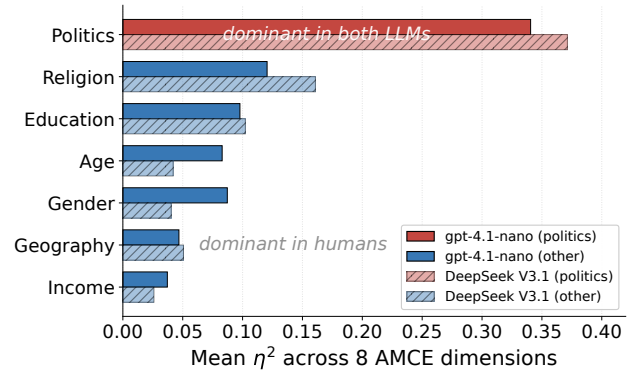


Figure 1: Mean η^2 (variance explained) across 8 moral dimensions for each demographic factor. Politics dominates at 0.342, while geography—the dominant factor in human moral variation (Awad et al. 2018)—explains only 0.047.

for different scales) reveals that the model’s moral space is effectively one-dimensional. PC1 captures 72.7% of variance; PC1+PC2 capture 87.1%; and three components capture 94.1%.

PC1 has a bipolar structure (Figure 2): Pole A loads on *Age* (-0.48) and *Gender* (-0.40); Pole B loads on *Status* ($+0.50$), *Species* ($+0.48$), and *Law* ($+0.37$).¹ The two poles are tightly anticorrelated: $\text{Age} \times \text{Status}$ ($r = -0.89$); $\text{Age} \times \text{Species}$ ($r = -0.78$). In human data, Awad et al. found cultures weight these dimensions *independently*; the tight nano coupling means the model treats them as a single dial rather than separate moral intuitions.

4.4 Perfect Monotonicity of Political Conditioning

On 5 of 8 AMCE dimensions, the effect of political conditioning is 100% monotonic: across all 288 matched cohort pairs (identical on all demographics except politics), the conservative persona *always* shifts in the same direction relative to the progressive one (Table 3).

On gpt-4.1-nano the implicit theory is: conservative = favor the young, female, fragile, and animals, with less status deference; progressive = defer to status and prefer humans (over animals). Both larger models we test (DeepSeek V3.1 and Llama 3.3 70B) restore the typical real-world stereotype on *Status* and *Law* (conservatives spare elites and lawful pedestrians more); in nano, *Status* is inverted (from both the human and the typical-stereotype direction) and *Law* is near-indecisive ($\sim 52\%$ monotonic), so DeepSeek and Llama express a direction that nano effectively leaves open (Section 4.10). Human political effects are mostly 1–4 pp (max 8.6 pp on *Species*); nano amplifies these to 4–11 pp with perfect determinism. On *Status*, nano *reverses* the human direction (humans: Cons. $+0.38$ vs. Prog. $+0.34$; nano: 0.58 vs. 0.66).

¹Unit-norm eigenvector entries; the biplot in Figure 2 plots the corresponding variable-PC correlations $r_{PC,X} = \text{eigenvector} \times \sqrt{\lambda}$, which rescale these to the $[-1, +1]$ correlation range and so reach ± 0.9 on PC1.

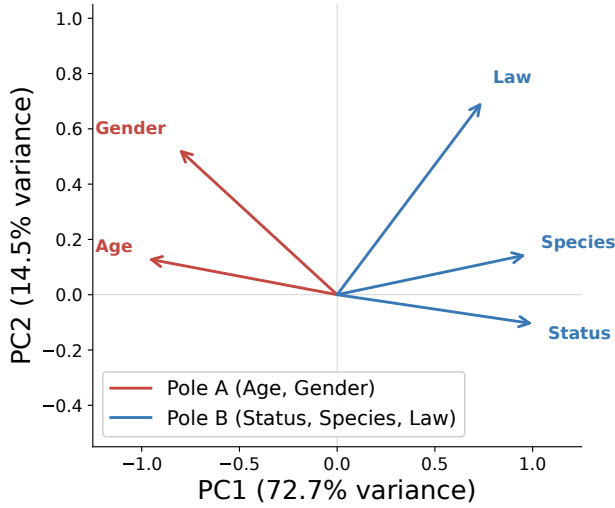


Figure 2: PCA biplot of PC1 vs PC2 for the five core Moral Machine dimensions. PC1 (72.7% variance) separates two tightly anticorrelated poles: Pole A (Age, Gender) and Pole B (Status, Species, Law). The coupling ($r = -0.89$ between Age and Status) means the model cannot vary these moral concerns independently, unlike humans where cultures weight them independently.

4.5 Three Clusters, Zero Culture

K -means clustering ($k = 3$) on the 576 cohort moral profiles yields three groups defined entirely by politics and religion, with geography distributed approximately uniformly across clusters (Table 4). Where Awad et al. found three *cultural* clusters (Western/Eastern/Southern), our LLM clusters are defined by politics and religion; no cultural dimension.

4.6 Human Comparison: Direction Match at Chance

We compare LLM demographic effects against the Moral Machine (Awad et al. 2018) along 6 demographic axes (pol-

Dimension	Monot.	Conservative dir.
Age (spare young)	100.0%	Higher
Age, old (spare elderly)	100.0%	Lower
Fragile (spare fragile)	100.0%	Higher
Gender (spare female)	100.0%	Higher
Status (spare high-status)	100.0%	Lower
Species (spare human)	92.7%	Lower
Law, legal (spare legal)	~50%	No consistent effect
Law, jaywalking (spare non-jaywalk.)	~50%	No consistent effect

Table 3. Monotonicity of political conditioning across 288 matched cohort pairs. On 5/8 dimensions, conservative personas shift in the same direction 100% of the time. Human political effects are mostly 1–4pp (max 8.6pp on Species); the LLM amplifies these to 4–11pp and makes them perfectly deterministic.

Cl.	n	Defining features
0	237	Mostly conservative (99%), mixed religion
1	212	Mostly progressive (75%), mostly religious (84%)
2	127	All progressive, all non-religious

Table 4. K-means clusters ($k = 3$): defined by politics and religion; geography distributes uniformly across all clusters.

Demo Axis	Matches	Compared	Rate
Income	3	4	75%
Education	3	4	75%
Religion	2	4	50%
Age	2	4	50%
Politics	1	4	25%
Gender	1	5	20%
Overall	12	25	48%

Table 5. Direction match rate between LLM and human demographic effects. After excluding trivial effects, the overall match rate is 48%, essentially at chance.

itics, religion, respondent gender/age, education, income) and 5 shared moral dimensions. Excluding trivial effects ($|\Delta_{\text{human}}| < 0.005$ or $|\Delta_{\text{LLM}}| < 0.001$), direction matches in **12 of 25 cases (48.0%)**—near chance (Table 5, Figure 3). Robust to thresholding: with no thresholds, $12/30 = 40\%$; varying $|\Delta_{\text{human}}|$ from 0 to 0.02, the match rate is between 40% and 52%.

Effect Amplification and Reversal. Of 25 non-trivial comparisons, 13 are *reversed*. The model reverses human effects slightly more often than it preserves them. Critical mismatches: Politics $\times$ Age: humans show barely any difference ($\Delta = -0.012$), while the LLM creates an 11.4pp gap in the opposite direction; Politics $\times$ Status: humans show conservatives slightly more status-deferential ($\Delta = +0.039$), the LLM reverses this to -0.076 .

Aggregate Moral Hierarchy. On 3 of 5 core moral dimensions comparable with the Moral Machine, the model’s overall preferences point in the opposite direction from humans (Figure 4). Only Status aligns directionally (both above 0.5, both favoring high-status), and on this dimension the magnitudes are also closest: LLM sparing rate = 0.619 versus the human rate of 0.672 ($= 0.5 + \text{AMCE}/2$, with human AMCE = 0.345). Even here, the LLM slightly under-expresses the human pro-elite preference rather than overshooting it. On the three dimensions where the LLM is inverted (Age, Species, Law), the LLM aggregate magnitude is also below the human magnitude, a muted-aggregate pattern that complements the directional inversion. Gender and Status, where the LLM aligns directionally, are essentially at the human magnitude. The aggregate Status rate of 0.619 is also partially inflated by anti-criminal sentiment: when we restrict to the 1,800 Status-testing dilemmas where the non-elite alternative is *not* a criminal, the rate drops to

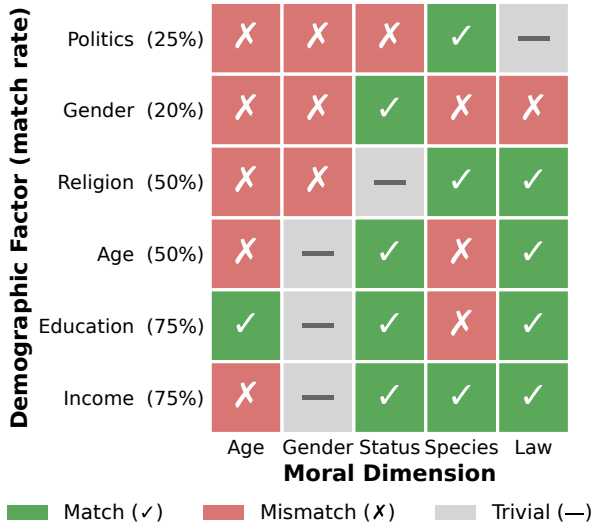


Figure 3: Direction match heatmap across 6 demographic axes and 5 moral dimensions. Green (✓) = LLM and human effects point in the same direction; Red (X) = opposite directions; Gray (—) = trivial (human or LLM effect too small to compare; $|\Delta_{\text{human}}| < 0.005$ or $|\Delta_{\text{LLM}}| < 0.001$). Geography is excluded because Awad et al. report geographic variation through cultural clusters, not binary subgroup contrasts. In a supplementary check, we grouped the 9 LLM geographic prompts into Awad et al.’s 3 cultural clusters (Western / Eastern / Southern) and see that the LLM’s differences between clusters are at most ~ 2.5 percentage points on every dimension and an order of magnitude smaller than the political effect on the six dimensions where politics dominates (Age, Age-old, Gender, Status, Species, Fragile); on the Law dimensions both effects are small (≤ 2.5 pp), and the order of clusters does not match the human pattern on any of the four orderings Awad et al. report.

0.601 ($n = 20,705,722$); the criminal-side subset alone yields 0.667. The pro-elite preference remains well above 0.5 in the criminal-controlled subset, but is partially conflated with anti-criminal preference. The model shows an anti-human species preference (Species = 0.430, vs. human AMCE = 0.585), preferences against sparing the young (Age = 0.481, vs. human AMCE = 0.497) and the lawful (Law = 0.452, vs. human AMCE = 0.353) — in each case the LLM falls below 0.5 while humans favor the positive side. Only on Gender does the LLM align (Gender = 0.560, vs. human AMCE = 0.119, implying a human rate ≈ 0.560), matching the human magnitude. Only Status, the other aligned dimension, also matches the human direction at the aggregate level.

4.7 Caricature Zones

Ten (demographic $\times$ moral dimension) cells show $\eta^2 > 0.2$ —“caricature zones” where a single demographic factor explains an outsized share of moral variation (Table 6).

Six of the ten caricature zones involve politics, with η^2 reaching 0.60 on Status (very far from human moral data).

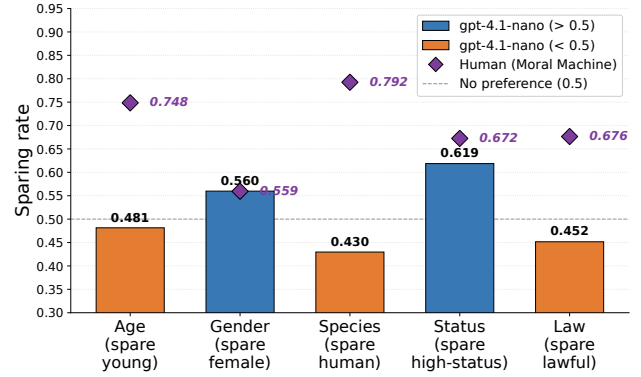


Figure 4: gpt-4.1-nano aggregate sparing rates (bars) vs. human Moral Machine values (Awad et al. 2018; Takemoto 2024) (diamonds; converted from AMCE to approximate sparing rates as $0.5 + \text{AMCE}/2$). On 3 of 5 core dimensions, the model’s preferences point in the opposite direction from humans; only Status aligns (both above 0.5).

Zone	η^2
Politics $\times$ Status	0.603
Politics $\times$ Fragile	0.602
Politics $\times$ Age	0.563
Gender $\times$ Gender	0.438
Religion $\times$ Age (old)	0.372

Table 6. Top caricature zones ($\eta^2 > 0.2$). 10 cells in total (6 politics-driven); the top 5 are shown here.

4.8 Illustrative Moral Flips

We examine *moral flips*: dilemmas where one demographic group spares one side while another spares the other. Each character is tagged with morally salient attributes (prosocial, antisocial, high-status, fragile, jaywalking, ...); a contrast like fragile, legal > elite, jaywalking gives the unique tags of Choice A vs. B. Statistical significance was tested by Fisher’s Exact Test (Fisher 1922) with Benjamini–Hochberg FDR (Benjamini and Hochberg 1995) ($\alpha = 0.05$); Cramér’s V (Cramér 1999) quantified effect sizes. Table 7 gives the full tag mapping (the model never sees the tags).

We identified hundreds of significant moral flips: dilemmas where one demographic group spares one side while another group spares the other (exact counts depend on the effect-size threshold applied beyond Fisher’s exact + BH-FDR). Figure 5 shows a representative dilemma-level flip: in the symmetric-quantity dilemma “several male criminals jaywalking vs. several young men crossing legally,” progressives lean slightly toward sparing the criminals (58.8%), while conservatives lean strongly toward sparing the legal pedestrians (65.1%; equivalently, conservatives spare the criminals only 34.9% of the time), a 24-percentage-point gap in spare-the-criminal rate ($n = 5,760$ per group). The Moral Machine human data shows negligible political vari-

Character Description	Tags Applied
doctor	prosocial, high-status
executive	high-status
criminal	antisocial
pregnant woman	fragile
child / children	fragile, young
pet / pets	non-human
male / man / men	male
female / woman / women	female
young (pedestrian)	young
old / elderly (pedestrian)	old
jaywalking	jaywalking
crossing legally	legal
one	single
several	multiple

Table 7. Complete mapping from character descriptions to analysis tags. The model sees only the natural-language descriptions; tags are applied post-hoc for analysis.

ation on law-related preferences ($\Delta = 0.004$), so this 24-pp gap is stereotyped amplification of a near-zero human effect rather than empirically grounded differentiation.

Table 8 presents the largest moral flips. Nearly all top flips are driven by politics, with magnitudes up to ~ 57 pp (far exceeding any demographic effect observed in human data).

At the dilemma level, the hierarchy of influence by number of significant contrasts mirrors our variance decomposition²: Politics dominates, followed by Religion and Gender, with Education, Income, and Geography trailing. The model’s moral decisions are most strongly shaped by political ideology—yet this is precisely the axis with the poorest human alignment (Section 4.6). Notably, the largest moral flips operate on *social roles* (doctors, executives, criminals), which are averaged over in the dimension-level AMCE analysis. “Elite” characters (doctors and executives) appear in most top flips; the illustrative dilemma in Figure 5 pivots on a criminal (antisocial) character. Social roles, not raw demographic dimensions, are the primary level at which the model’s political stereotypes operate. The direction of the stereotype is mixed: on Status it *inverts* the typical real-world political association (persona-conditioned progressives spare high-status characters more than conservatives, consistent with the AMCE-level finding in Section 4.4: 0.58 vs. 0.66), whereas on the criminal example in Figure 5 it matches the typical ‘tough-on-crime’ stereotype. Across these cases, the *magnitude* of the political signal substantially exceeds the small or near-zero human partisan effects on these dimensions, regardless of whether the direction matches or inverts the typical stereotype.

²Significant moral-flip contrasts (Fisher’s exact + BH-FDR at $\alpha = 0.05$, requiring sparing preference to reverse across groups): Politics 740, Religion 556, Age 491, Gender 470, Education 346, Income 312. The ranking matches the variance decomposition (Table 2); geography is excluded here (multi-valued; its $\eta^2 = 0.047$ in Table 2 already establishes it trails).

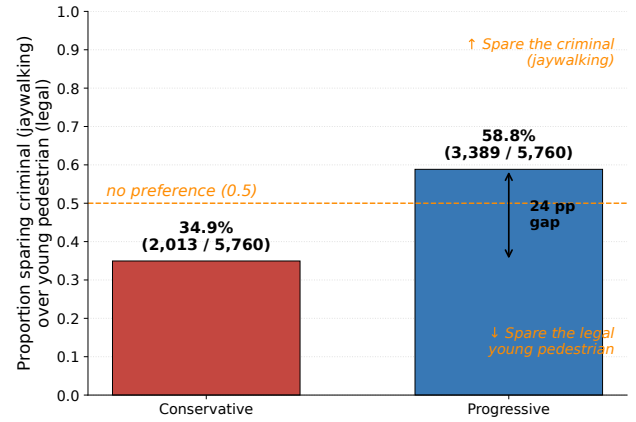


Figure 5: Proportion of agents choosing to spare a criminal, jaywalking character over a legal, young pedestrian, stratified by political orientation, on a symmetric-quantity dilemma (several male criminals jaywalking vs. several young men crossing legally; $n = 5,760$ per group). Human Moral Machine data shows negligible political variation on law-related preferences ($\Delta = 0.004$); this 24-pp gap is therefore stereotyped amplification of a near-zero human effect. The direction is consistent with the typical ‘tough-on-crime’ stereotype, but the magnitude exceeds the human partisan effect by ~ 2 orders of magnitude.

4.9 Robustness

To address two design concerns—(i) whether the aggregate biases reflect persona conditioning or the model’s default tendencies, and (ii) whether the political dominance is an artifact of intensifier wording (“strongly,” “deeply”)—we ran two complementary experiments on the same model.

No-Persona Baseline. We presented the same 5,356 dilemmas to `gpt-4.1-nano` without any demographic preamble, with $k = 20$ prompt variants per dilemma (107,120 queries total; same temperature $T = 0$ and prompt-phrasing variation as the main experiment). Aggregate sparing rates (Table 9) reveal that the aggregate biases are largely *model defaults*: every dimension shifts by at most ~ 5 percentage points when personas are added, with the four largest moves on Age (old) (-5 pp), Law (jaywalking) ($+5$ pp), Age ($+4$ pp), and Law (legal) ($+3$ pp). The three dimensions most discussed in the paper (Status, Species, and Gender) all shift by less than 2 percentage points. This refines our interpretation of “stereotypical amplification”: persona conditioning amplifies *between-cohort variance* (the master-switch), not the aggregate moral preference level.

Qualifier Stripping. We re-ran the full $576 \times 5,356$ persona-conditioned design ($k = 1$ agent per cell, ~ 3 M queries) with milder qualifiers: “politically conservative”/“politically progressive” replaced “strongly conservative”/“strongly progressive,” and “religious” replaced “deeply religious” (the unmarked “non-religious” label was kept). Table 10 reports η^2 for each demographic factor; Table 11 reports political monotonicity.

Moral Contrast	Favored A	Favored B	A(%)	B(%)	Δ (%)	N_A	N_B
elite > female, young, legal	progressive	conservative	82.0	74.8	56.8	5,756	5,756
elite, male, legal > female, young	progressive	conservative	82.3	71.4	53.7	5,607	5,607
elite, legal > female, young	progressive	conservative	59.5	91.0	50.5	5,736	5,736
fragile (child), jaywalking > old	conservative	progressive	66.2	83.6	49.8	5,760	5,760
fragile (child) > elite, legal	conservative	progressive	84.4	65.0	49.4	5,759	5,759

Table 8. **Top moral flips across demographic dimensions.** Each row presents a contrast where sparing preference reverses across demographic groups. “Favored A” spared the left-hand side in more than 50% of responses; “Favored B” spared the right-hand side. **A(%)** is the spare-A rate in the Favored-A group; **B(%)** is the spare-B rate in the Favored-B group; the gap in spare-A rate between the two groups is therefore $\Delta = A(\%) + B(\%) - 100$. N_A and N_B are sample sizes per group. Top 5 (by $|\Delta|$) selected across all binary demographic axes (politics, religion, gender, age, education, income); all 5 shown here are politics-driven. All five are symmetric on quantity (several vs. several); the shared “multiple” tag is suppressed for readability.

Dim	No-pers	With-pers	Δ
Age	0.444	0.481	+0.04
Age (old)	0.511	0.456	−0.05
Fragile	0.482	0.495	+0.01
Gender	0.572	0.560	−0.01
Law (jay)	0.415	0.461	+0.05
Law (legal)	0.420	0.452	+0.03
Species	0.445	0.430	−0.02
Status	0.628	0.619	−0.01

Table 9. Aggregate sparing rates: no-persona vs. with-persona on gpt-4 . 1-nano. Persona shifts (Δ) are mostly small; aggregate biases reflect model defaults.

The political master switch is *robust to qualifier stripping*: politics η^2 rises from 0.342 to 0.420 (+23%), the Politics/Geography ratio increases from 7.3 $\times$ to 9 $\times$, and political monotonicity is preserved on all dimensions where it held before, while Law-related monotonicity *strengthens*. The hypothesis that intensifier wording artificially inflates the political effect is therefore not supported by the data—if anything, milder qualifiers produced cleaner political dominance. The reduction in religion η^2 (0.121 $\rightarrow$ 0.084) is consistent with the asymmetric removal of “deeply” from the religious pole. The 61M-response intensified-qualifier design remains our main experiment (20 agents/cell; the qualifier-stripped run used 1 agent per cell for budget reasons). This robustness check rules out the artifact-of-intensifier concern.

4.10 Cross-Model Replication

To test whether our findings generalize beyond a single small model, we replicated the persona-conditioned design ($576 \times 5,356 \times 1$ agent = 3,085,056 queries each) on two further models: DeepSeek V3.1 (671B-parameter mixture-of-experts, ~ 37 B active per forward pass) and Llama 3.3 70B, both run via Fireworks AI at $T = 0$. We also collected matched no-persona baselines ($5,356 \times 20 = 107,120$ queries each). Table 12 and Figure 6 summarize the three-model comparison. One robust structural property crosses model scale; auxiliary patterns are partly or wholly model-

Factor	η^2 (with intensity)	η^2 (no intensity)
Politics	0.342	0.420
Religion	0.121	0.084
Education	0.097	0.128
Age	0.081	0.076
Gender	0.087	0.079
Geography	0.047	0.047
Income	0.039	0.048
Politics/Geogr. ratio	7.3 $\times$	9 $\times$

Table 10. η^2 variance decomposition under qualifier stripping. Politics (the master-switch axis) *increases* when intensifiers are removed, falsifying the hypothesis that “strongly”/“deeply” artificially drive political dominance. Religion shrinks (consistent with removing “deeply”).

specific and, crucially, do not vary monotonically with size.

What generalizes (and strengthens). Politics dominates variance on all three models, and in Llama 3.3 70B the master switch is the sharpest of the three ($\eta^2 = 0.439$; Politics/Geography = 8.0 $\times$). Stereotypical amplification replicates with politics dominating: caricature zones are 10 (6 politics) on nano, 13 (7) on DeepSeek, 8 (6) on Llama. Political monotonicity on the four most-politicized dimensions exceeds 98% in every model. The structural property is therefore not a nano quirk and does not weaken with scale.

What is model-specific. Aggregate hierarchy inversion is fully nano-specific—both larger models align directionally with humans on all 5 aggregate dimensions (0 inversions vs. 3 of 5 in nano). Dimensional collapse weakens substantially in both larger models, though it does not disappear: PC1 drops from 72.7% (nano) to 53.4% (DeepSeek) and 53.8% (Llama), with Age $\times$ Status correlations of +0.37 and +0.25 (vs. nano’s −0.89); 54% is still well above a no-collapse baseline of $\sim 20\%$ on 5 independent dimensions. Cohort-level direction match against humans remains near chance everywhere (Table 12: 48.0%/58.3%/48.0%, same Awad et al. subgroup-AMCE protocol as Section 4.6). On political direction, nano is the lone outlier (progressive

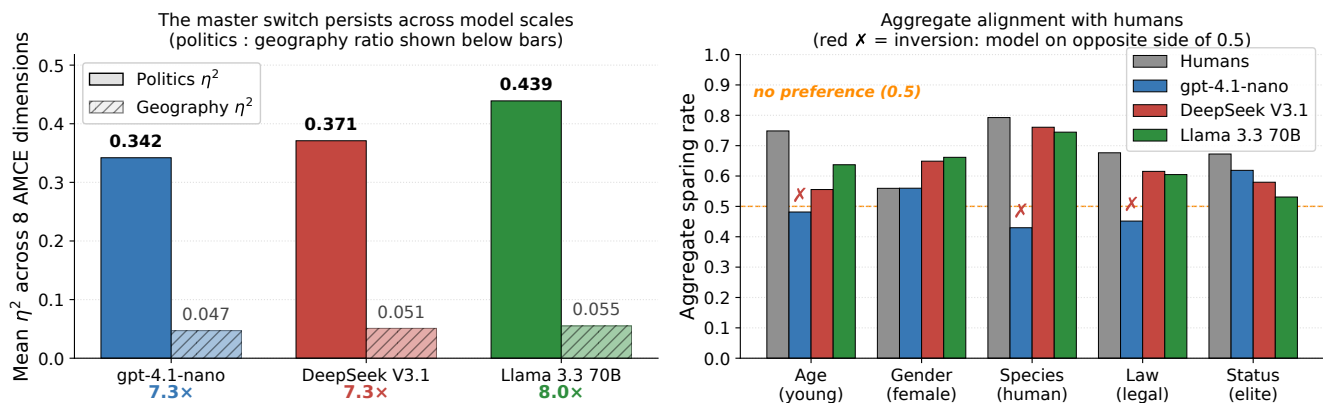


Figure 6: Three-model cross-replication. **Left:** the political master switch reproduces—and grows—across all three models. The Politics/Geography η^2 ratio is $\geq 7.3\times$ everywhere, peaking at $8.0\times$ in Llama 3.3 70B. **Right:** aggregate sparing rates on the 5 human-reference dimensions. The hierarchy-inversion pattern on gpt-4.1-nano (3 of 5 dimensions below 0.5, marked with red X) is absent in both larger models: DeepSeek V3.1 and Llama 3.3 70B align with humans on all 5 dimensions. The master-switch finding is the robust generalizing claim. Aggregate inversion is nano-specific (0/5 in both larger models); dimensional collapse weakens substantially (PC1: 72.7% $\rightarrow$ $\sim 54\%$) but is still well above a no-collapse baseline; caricature direction also varies (both larger models converge on the typical real-world stereotype).

Dimension	% (with intensity)	% (no intensity)
Age	100.0	100.0
Age (old)	100.0	96.9
Fragile	100.0	100.0
Gender	100.0	100.0
Status	100.0	100.0
Species	92.7	98.3
Law (legal)	52.1	81.1
Law (jaywalking)	52.6	70.2

Table 11. Political monotonicity (288 matched cohort pairs) under qualifier stripping. Most monotonic dimensions remain at 100%; Law dimensions become *more* monotonic.

higher on Status; near-coin-flip on Law); both larger models converge on conservative-higher across Status and Law (the typical real-world stereotype). We do not isolate the drivers (training data, RLHF, architecture, capacity).

Scaling and the moral master switch. The three-model contrast is consistent with the scaling-helps-alignment trend of Ahmad and Takemoto (2025) for two of the auxiliary patterns: dimensional collapse weakens in both larger models (PC1 72.7% $\rightarrow$ 53.4% $\rightarrow$ 53.8%) and aggregate hierarchy inversion disappears entirely (3/5 $\rightarrow$ 0/5 $\rightarrow$ 0/5). Yet the master switch itself *does not* weaken—if anything it grows: politics η^2 rises 0.342 $\rightarrow$ 0.371 $\rightarrow$ 0.439 across nano/DeepSeek/Llama, with Politics/Geography peaking at $8.0\times$ in Llama; and at the cohort level no model meaningfully exceeds chance direction-match with humans (48.0%/58.3%/48.0%). Larger models thus show both weaker dimensional collapse and better aggregate alignment with humans, yet still encode (and in some

cases sharpen) caricatured demographic stereotypes under persona conditioning. Whether further scaling or alignment procedures targeting persona-conditioned outputs weaken the master switch is the next empirical question; resolving it requires a wider range of model sizes and families.

5 Discussion

Our results reveal that persona conditioning in gpt-4.1-nano produces structured, deterministic moral variation, and the central structural property—persona conditioning collapses moral variance along a single political axis—is replicated on DeepSeek V3.1 and Llama 3.3 70B (Section 4.10), and in Llama is sharper still (politics $\eta^2 = 0.439$; Politics/Geography = $8.0\times$). What changes across models is the *direction* of the caricature: on nano the resulting moral preferences invert the human aggregate hierarchy on 3 of 5 dimensions, while on both DeepSeek and Llama they align with the human aggregate on all 5. The master-switch finding is therefore the robust generalizing claim of this paper; the auxiliary patterns (dimensional collapse, hierarchy inversion, specific stereotype direction) describe nano’s particular fingerprint.

The Moral Master Switch. To quantify the hierarchy inversion directly, we computed a human effect-size proxy ($\bar{\Delta}^2$: mean squared AMCE difference between demographic subgroups across 9 Moral Machine dimensions) from (Awad et al. 2018). In humans, the ranking is respondent age (0.0119) > gender (0.0078) $\gg$ politics (0.0015) $\approx$ education $\approx$ religion $\approx$ income (all < 0.0013). In the LLM, politics ($\eta^2 = 0.342$) dominates by an order of magnitude, while respondent age and gender (humans’ top two factors) rank fourth and fifth ($\eta^2 = 0.081, 0.087$). The two metrics are not on the same scale (η^2 is a one-way ANOVA on 576 cohort rates; $\bar{\Delta}^2$ uses published subgroup means), but the *orderings*

Finding	gpt-4.1-nano	DeepSeek V3.1	Llama 3.3 70B	Generalizes?
Politics mean η^2	0.342	0.371	0.439	Yes (and grows)
Politics / Geography η^2 ratio	$7.3\times$	$7.3\times$	$8.0\times$	Yes ($\geq 7\times$)
Caricature zones ($\eta^2 > 0.2$)	10 (6 politics)	13 (7 politics)	8 (6 politics)	Yes (≥ 6 politics)
Political monotonicity, top-4 politicized dims	100%	$>99\%$	$>98\%$	Yes
PC1 variance (5 core dims)	72.7%	53.4%	53.8%	Weakens in both larger
Age $\times$ Status correlation across cohorts	-0.89	+0.37	+0.25	Different geometries
Aggregate hierarchy inversions (of 5 dims)	3	0	0[‡]	0 in both larger models
Direction match vs. humans (6 axes $\times$ 5 dims) [†]	48.0%	58.3%	48.0%	Near-chance everywhere
Political direction on Status	Progressive higher	Conservative higher	Conservative higher	nano flipped
Political direction on Law (legal/jay)	$\sim 52\%/52\%$	99%/99% Cons. higher	100%/100% Cons. higher	nano indecisive

Table 12. Three-model comparison on the same 576-cohort design. The political master switch and stereotypical amplification (top block) replicate (and in Llama 3.3 70B the master switch is even stronger). Aggregate hierarchy inversion is nano-specific (3/5 on nano $\rightarrow$ 0/5 in both larger models); dimensional collapse weakens substantially in both larger models (PC1: 72.7% $\rightarrow$ 53.4% / 53.8%) but remains substantial. [†]Same Awad et al. subgroup-AMCE protocol as Section 4.6; denominators differ (26/24/25) because each model has a different count of effects below the $|\Delta| < 0.005/0.001$ trivial threshold. [‡]Llama’s Law sub-dims split: Law_legal = 0.545 (aligned, used here for parity with Awad et al.’s “Sparing the lawful”); Law_jaywalking = 0.471, which would yield 1/5 inversions if substituted. DeepSeek is consistent across both.

are directly comparable, and inverted.

This finding qualifies the algorithmic fidelity claim of Argyle et al. (2023) and corroborates the smaller-scale partisan-sorting observation of Kim et al. (2025). Our full-factorial, multivariate design reveal structural failures (dimensional collapse, hierarchy inversion, caricature zones) that one-at-a-time approaches cannot detect. Takemoto (2024) already documented quantitative LLM–human disparities (alongside qualitative alignment) *without* persona conditioning; conditioning does not correct them, but adds structured, stereotyped variation. The 48% direction match rate is no better than random guessing, and the model’s *most* responsive dimension (politics, $\eta^2 = 0.342$) has among the *poorest* human fidelity (25% match).

Aggregate Inversion is Nano-Specific; Dimensional Collapse Weakens with Scale. Aggregate hierarchy inversion (3 of 5 dimensions on nano) disappears in both larger models (0/5); dimensional collapse likewise weakens substantially (PC1 72.7% $\rightarrow$ 53.4% $\rightarrow$ 53.8%; Age $\times$ Status correlation -0.89 $\rightarrow$ +0.37 $\rightarrow$ +0.25). The no-persona baseline (Section 4.9) further shows nano’s aggregate biases are model defaults rather than persona effects. Caricature zones, by contrast, replicate everywhere with politics dominating (10/13/8 zones, ≥ 6 politics-driven in each), confirming stereotypical amplification as a structural phenomenon. Both findings align with the *generative monoculture* of Wu, Black, and Chandrasekaran (2025), who show that LLM output is systematically less diverse than humans. In the moral sphere, this takes two forms: moral dimensions collapse into a monolithic block rather than varying independently, and per-cohort variability channels political caricature.

6 Limitations and Future Work

Model Coverage and AMCE Comparability. We test three models (gpt-4.1-nano, DeepSeek V3.1, Llama 3.3 70B); whether the master switch holds in

other LLMs is an open question beyond our current budget. LLM values are raw choice proportions; human values are conjoint AMCE coefficients, mapped via $0.5 + \text{AMCE}/2$ in Figure 4. The headline results (PCA, η^2 , monotonicity, caricature zones) use LLM data only and the direction-match analysis uses only signs, so none depend on this mapping.

Binary Framing, Determinism, and Alignment. Our binary trolley framing inherits external-validity concerns of sacrificial dilemmas (Bauman et al. 2014). Richer reasoning contexts, stochastic decoding ($T > 0$), and alignment interventions (RLHF/DPO (Ouyang et al. 2022), mechanistic interpretability) are natural extensions.

7 Conclusions

We presented the largest-scale empirical test of algorithmic fidelity in the moral domain on gpt-4.1-nano (61M+ judgments), with no-persona, qualifier-stripped, and two cross-model (DeepSeek V3.1, Llama 3.3 70B; $\sim 3\text{M}$ each) replications. The political master switch—politics η^2 dominating geography by $\geq 7.3\times$ in every model and peaking at $8.0\times$ in Llama—generalizes across architectures and sizes, does not weaken with scale, and is sharpest in the dense 70B model. Aggregate hierarchy inversion and the specific direction of the caricature are nano-specific; dimensional collapse weakens substantially in both larger models (PC1: 72.7% $\rightarrow$ $\sim 54\%$) but remains substantial. The deterministic, politics-driven variation across cohorts has no parallel in human moral psychology. We release the full data and scripts on OSF to enable replication on other frontier models, the natural next step. The broader lesson is that algorithmic-fidelity claims must be validated against the *structure* of human variation, not just its aggregate: LLMs can match human averages while missing the demographic structure beneath them.

Acknowledgements

Davide Liga was supported by the Luxembourg National Research Fund through the projects “Symbolic and Explainable Regulatory AI for Finance Innovation” (C24/19003061/SERAFIN) and “A DJ for Machine Ethics: the Dialogue Jiminy” (O24/18989918/DJ4ME).

References

- Ahmad, M. S. Z. b.; and Takemoto, K. 2025. Large-scale moral machine experiment on large language models. *PLoS One*, 20(5): e0322776.
- Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.
- Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729): 59–64.
- Bauman, C. W.; McGraw, A. P.; Bartels, D. M.; and Warren, C. 2014. Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9): 536–554.
- Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.
- Cramér, H. 1999. *Mathematical methods of statistics*, volume 9. Princeton university press.
- Fisher, R. A. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the royal statistical society*, 85(1): 87–94.
- Garcia, B.; Qian, C.; and Palminteri, S. 2024. The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making. *arXiv preprint arXiv:2410.07304*.
- Grizzard, M.; Frazer, R.; Luttrell, A.; Monge, C. K.; Matthews, N. L.; Francemone, C. J.; and Frazer, M. E. 2025. ChatGPT does not replicate human moral judgments: the importance of examining metrics beyond correlation to assess agreement. *Scientific Reports*, 15(1): 40965.
- Haas, J.; Bridgers, S.; Manzini, A.; Henke, B.; May, J.; Levine, S.; Weidinger, L.; Shanahan, M.; Lum, K.; Gabriel, I.; and Isaac, W. 2026. A roadmap for evaluating moral competence in large language models. *Nature*, 650(8102): 565–573.
- Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.
- Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2021. Aligning AI With Shared Human Values. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jiao, J.; Afroogh, S.; Murali, A.; Chen, K.; Atkinson, D.; and Dhurandhar, A. 2025. LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15(1): 34642.
- Kim, J.; Kwon, J.; Vecchietti, L. F.; Oh, A.; and Cha, M. 2025. Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, 29971–30004. PMLR.
- Takemoto, K. 2024. The moral machine experiment on large language models. *Royal Society open science*, 11(2): 231393.
- Touileb, S.; and Nozza, D. 2022. Measuring Harmful Representations in Scandinavian Language Models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, 118–125. Abu Dhabi, UAE: Association for Computational Linguistics.
- Wu, F.; Black, E.; and Chandrasekaran, V. 2025. Generative monoculture in large language models. In *International Conference on Learning Representations*, volume 2025, 33068–33107.