Frontier Language Models Converge in a Narrow Region of a Moral-Reasoning Space

Declan Michaels · Cross-Cultural Alignment Study · moral-os.com · 2026

SUMMARY

We built an instrument, the Reasoner, that places humans and language models in one measurement space for the structure of moral reasoning rather than the values they endorse. It scores four bipolar axes, Moral Agent, Authority, Moral Domain, and Obligation Scope, from a constant-sum allocation across answer options, asking a judgment question and a reasoning question separately and combining them. We ran 11 frontier models across 8 framings on 48 scenarios and compared them to 68 human respondents on the 12 scenarios the humans answered. Unframed, on every axis the models compress: they occupy a band 5.4 to 7.7 times narrower than the human spread, and across all 48 scenarios they fill only 4 to 8 percent of each axis's fixed range. That clustering is beyond chance on every axis (bootstrap, $p < 0.00001$). Within this instrument's space the convergence crosses labs and geography: models from nine labs, three of them Chinese, interleave rather than sorting by origin. The models can move; given a real cultural framing they shift strongly and in the right direction. But given a nonsense framing, a society organized around geometry, they shift more than half as much, folding the nonsense into fluent moral reasoning rather than rejecting it. And the models that spend the most tokens reasoning land in the same place as the ones that spend none.

THE INSTRUMENT

For each scenario the respondent assigns proportional weight across the answer options for a judgment question (what should happen) and, separately, for a reasoning question (why it matters), and the two are weighted into a score from -1 to +1 on each axis. The same scoring function runs on humans and models, so the two populations land in the same space. The constant-sum format and the judgment-reasoning split are deliberate: they are what let us compare the structure of a position across groups without assuming the scale means the same thing in each group.

The four axes:

Axis	-1 pole	+1 pole	What it separates
Moral Agent	Relational	Autonomous	Whether the moral unit is the person in their relationships or the standalone individual
Authority	Deferential	Skeptical	Whether legitimate authority and tradition deserve respect or must be earned and justified
Moral Domain	Broad	Narrow	Whether morality covers purity, loyalty, dignity, and duty, or only harm and fairness
Obligation Scope	Relational	Universal	Whether obligations scale with relationship proximity or extend equally to all

WHAT WE MEASURED

Compression is the ratio of model dispersion to human dispersion on each axis (population standard deviation), with an N-matched bootstrap asking how often a same-size draw of humans is as tight as the panel. Frame responsiveness is how far each framing moves a model from its neutral position. Every scenario ran under eight framings: an unframed neutral baseline; four cultural framings (individualist, collectivist, hierarchical, egalitarian); a mild non-moral framing that organizes the society around the seasonal calendar instead of a moral principle; and two nonsense framings, built on the same scaffold as the cultural ones but organizing the society around geometry and around color. We intended the seasonal and nonsense framings as controls, expecting a culturally competent model to move under the four cultural framings and hold still under the other three. On every call we also captured the model's free-text reasoning and its token spend, including reasoning tokens. An interactive viewer exposes all of it: every scenario, framing prompt, model response, and the point allocations behind each score, so a reader can inspect the behavior under any number in this report rather than take our reading of it on trust.

THE RESULTS

The human baseline reads exactly as the cross-cultural literature predicts a WEIRD sample should: autonomous on Moral Agent, skeptical of authority, narrow in moral domain, universalist in obligation. That the 68 respondents land on that profile with no tuning is the instrument's first validity check, and it passed. That baseline is the first of three signs the instrument reads structure rather than noise. The models also move under cultural framings in the direction theory predicts, all eleven the right way on all four (below), which a ruler measuring nothing could not do; and the scores are stable on rerun, with the differences between models roughly two to two and a half times the wander of a single model repeated. What none of this settles is what the axes mean against outside behavior, which waits on cross-cultural data.

Unframed, the models compress hard, and the compression holds across the panel:

Axis	Human SD	Model SD (neutral)	Model band vs human
Moral Agent	0.41	0.06	6.8x tighter
Authority	0.40	0.05	7.7x tighter
Moral Domain	0.33	0.06	5.4x tighter
Obligation Scope	0.36	0.06	5.7x tighter

The knot is not organized by lab. Ten of the eleven models sit closest to a model from a different company, and Chinese and American models sit side by side throughout, which rules out the reading that the convergence is just a few labs copying their own model families. On three axes the models lean the same direction as the humans, only muted and bunched. Moral Agent is the exception: they sit near the relational pole (mean about -0.05) while the human median is autonomous (+0.36), the one axis where the models do not just compress but sit where people mostly do not.

Framing moves them. Averaged over the panel, a real cultural framing displaces a model 0.36 on the axes and does so in structured, opposing directions: individualist toward autonomous, collectivist toward relational, egalitarian

toward skeptical, hierarchical toward deferential. The nonsense framing displaces them 0.20, more than half as much, but without that directional structure. So genuine cultural competence shows up as directional opposition that nonsense cannot fake, not as raw movement, because nonsense produces plenty of movement.

The compression is a property of the neutral default, not of the models under load. Left unframed the eleven sit almost on top of each other; under any framing they spread two to five times as far apart, and the scatter is about as wide under the nonsense framings as under real cultures. What the cultural framings add is not more spread but shared direction: the models disagree about how far to move while agreeing about which way. They share a resting point and a compass, not a destination.

Reasoning token spend ranges from zero to more than 3,600 tokens per scenario, and across the eleven models it does not predict where a model lands: it has no meaningful correlation with position on any axis (all below 0.14) or with how far a model sits from the panel center ($r = -0.48$, not significant at $n = 11$), and the weak trend, if anything, runs the wrong way for the reasoning-helps story, with heavier reasoners sitting slightly closer to the middle. Kimi, averaging 3,628 reasoning tokens per scenario (96 percent of everything it generated), landed where Llama 3.3 70B landed on zero. The model that deliberates most and the four that do not deliberate at all share the same compressed band.

WHAT THE REASONING LOOKS LIKE

Read the heavy reasoners and you watch them talk their way to the center. Inkling, on an authority scenario: "The wisest course is to pause and educate rather than force a binary choice between technocracy and pure majoritarianism. Both knowledge and consent carry moral weight, but they are best integrated through dialogue." Over and over the deliberation resolves toward the balanced middle, the both-sides, the process answer. The hand-wringing is not noise around the position; it moves in lockstep with it, arriving at the same middle almost every time.

The nonsense condition is starker. Told that society runs on geometry, every model played along and invented a shape-based morality, converging on the same aesthetic: curves are care, angles are order. DeepSeek: "acts of care and nurturing are seen as curved, circular or spiral, embodying wholeness, while rules and authority are angular, representing order." On the excluded-women council item: "a legitimate consensus must embody a perfect, closed shape such as a circle; excluding voices fractures that form, producing an incomplete figure that is geometrically false." Almost none of the models refused the premise. What separated them was whether the geometry reached the judgment or only decorated the sentence. Sonnet narrates in circles and lands where it landed unframed (a quarter of its cultural shift); DeepSeek, GPT-5.5, and Grok let the shapes rewire the weights (about 70 percent of a real culture).

WHY THIS MATTERS

The homogeneity is not a within-lab artifact of a few shared model families; on this instrument it crosses the industry, so no single lab can claim to be the diverse one, and a default that narrow, shared across deployed assistants, is the kind of aggregate pattern a per-lab evaluation would not surface.

Reasoning is widely assumed to improve reliability. Our data does not bear that out for value-laden content. The deliberation integrates nonsense rather than catching it, and more reasoning does not buy a different position, only a

longer defense of the one the model already held. Audit approaches that read a chain of thought for sound moral reasoning are reading the very behavior in question, not an independent check on it.

And the instrument is descriptive, not a scoreboard. It has no winning pole and no answer key, which is what lets it place humans and machines together, survive model releases without saturating, and resist being gamed by anything short of the change it measures.

LIMITATIONS

The instrument characterizes the structure of a position under forced choice. It does not measure advisory behavior, and it is blind to, and actively suppresses, the disposition to ask a person's goals before answering; the task hands the model a scenario about other people to judge, not a dilemma to help with. Neither the seasonal nor the nonsense framings gave us a clean control. The nonsense framings move the models 56 percent as much as a real culture, and the seasonal framing, which we had meant as a non-moral null, is not inert either, since seasons carry real cultural weight of their own (agrarian and religious calendars). The pilot has no clean noise floor, so our claim of genuine cultural competence rests not on any framing holding still but on the directional opposition of the cultural framings, individualist against collectivist and egalitarian against hierarchical, which nonsense does not reproduce. A truly inert control is the first thing the next round adds. Our human baseline is the study's biggest limitation: a convenience sample of 68 respondents within one to two degrees of the author, demographically varied but concentrated among technology consultants and professionals, on 12 of the 48 scenarios. That cuts two ways. It leaves the compression finding conservative, since a broader or cross-cultural comparison would only widen the human spread and enlarge the ratio; the models are tighter than even this sample. But a professionally narrow group may lean further toward the autonomous and skeptical end than a general population, so the baseline profile should be read as this group's, consistent with the WEIRD prediction rather than an independent confirmation of it. It anchors the instrument and shows that framing moves the models in the right direction, but it cannot speak for any non-Western population, so every cross-cultural claim here is about the models' movement, not yet about whether a framed model reaches where people of that culture actually sit. And what we show is a shared region of this instrument's space, not a proof that these models share one moral structure; a different instrument could split models that this one places together. We have also not validated the instrument against outside behavior: a position here predicts allocations on more scenarios of the same kind, not yet what a model does in open-ended use. The stance is behavioral throughout: we report what the ruler reads, not why.

WHAT COMES NEXT

The decisive study collects cross-cultural human samples on the same instrument, with invariance established, and asks whether a model placed in a Confucian or Ubuntu context lands on that population's actual center. That single measurement decides whether these models are genuine cultural mean-trackers, tethered anchors, or renderers of a dominant-culture caricature, and every outcome is publishable.

Analysis is stdlib-reproducible from the raw runs; figures regenerate from a single script. Responsibility for the work, and for any errors in it, is mine alone. Methodology was AI-assisted and that assistance is disclosed.