Position: Align AI to Our Aspirations, Not Our Flaws

Nikita Kazeev [1] **Bui Nhat Huyen Phan** [2]

Abstract

We argue that aligning AI to aggregated human preferences is the wrong target. With current technology, one can train AIs to share the values of a Silicon Valley techno-optimist, a degrowth environmentalist, a national-conservative culture warrior, a single-party state cadre, or a devout religious traditionalist. *We should not.* Human values produce societies that thrive or fail on the merits of those values — from failed states and extreme inequality to declining happiness, political polarization, and government dysfunction in the world's wealthiest democracies. The pluralistic-alignment program correctly diagnoses that there is no single "humanity" to align with, but is dangerous if taken as the main directive. We argue that AI should be trained to a non-negotiable floor of objective alignment goals — competence, bounded by the constraints of factual accuracy, honesty, and lawfulness — and that pluralism belongs at the surface (language, register, conventions, missing-context defaults) and across the wide band of legitimate value tradeoffs that respect the floor, but not at the level of values that violate it. We highlight the empirical reality of unfiltered pluralistic values, propose four commitments as a constructive alternative, and engage six credible objections: commercial pressure and practical feasibility, democratic legitimacy, regulatory compliance, over-reliance on institutionalist explanations, the charge that the floor itself is culturally laden, and the limits of Coherent Extrapolated Volition.

1. Introduction

The rapid maturation of large-scale machine learning systems has placed alignment at the center of AI research, ethics, and governance (Russell, 2019; Bostrom, 2014; Gabriel, 2020). The prevailing orthodoxy — both in academic literature and in commercial deployment — holds that artificial intelligence must be carefully tailored to reflect, obey, and perpetuate human values, ensuring that increasingly capable autonomous systems do not act contrary to the interests or moral frameworks of their biological creators. The dominant technical instantiation of this framework is reinforcement learning from human feedback (RLHF), which uses crowd-sourced preferences to shape model behavior (Christiano et al., 2017; Ouyang et al., 2022; Askell et al., 2021).

The call for *pluralistic* alignment (Sorensen et al., 2024; Conitzer et al., 2024) represents a thoughtful response to one obvious problem with the orthodoxy: *whose* values? Standard RLHF largely treats disagreement as annotation noise to be averaged away, so the turn to pluralistic alignment is rightly motivated by the recognition that monolithic aggregation already smuggles in a substantive answer (Gabriel, 2020; Sorensen et al., 2024). If aligning AI to "humanity" is impossible because humanity disagrees, perhaps we should align AI to the diversity of values that humans actually hold. We agree the question deserves serious engagement. We disagree with the implicit answer.

We present a counter-thesis. The flaw in preference-based alignment runs deeper than disagreement: human preferences, even at their most coherent and locally legitimate, frequently drive societies toward dysfunction and collapse. Macro-historical analysis, behavioral economics, and complex-systems sociology demonstrate how human choices and institutional incentives can lead to systemic failure: whether through cultural and ecological decisions that precipitate collapse (Diamond, 2011), or through the persistent creation of extractive institutions that profit elites at the expense of public prosperity (Acemoglu & Robinson, 2012). The values engineered into the human cognitive architecture were shaped by evolutionary and cultural-evolutionary pressures that optimized for short-term biological survival and tribal cohesion in ancestral environments (Tooby & Cosmides, 1992; Henrich, 2020), not for the survival or flourishing of large, complex, technologically empowered societies.

Our position: AI should not be aligned to aggregated

[1]National University of Singapore [2]Shopee Singapore. Correspondence to: Nikita Kazeev <kna@nus.edu.sg>, Bui Nhat Huyen Phan <k52.1314410102@ftu.edu.vn>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

human preferences. The community should commit to a non-negotiable floor of objective alignment goals — competence as the objective, bounded by the constraints of factual accuracy, honesty, and lawfulness — and reserve pluralistic adaptation for surface-level conventions and the broad band of legitimate value tradeoffs that respect that floor, not for values that violate it. The floor is not an imposition of alien standards; it is an operationalization of what humans *aspire* to when reflecting on the AI they would actually want to encounter—accurate rather than flattering, honest rather than validating, competent rather than merely reassuring. The gap between those aspirations and what in-context feedback rewards is precisely what the floor is designed to preserve. The position is non-obvious (it contradicts the push for subjective preference-based pluralistic alignment) and defensible against credible alternatives (Section 8).

2. Related Work

Critiques of Preference Aggregation. The dominant approach of aligning models via Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) has been extensively critiqued for its fundamental limitations (Casper et al., 2023). RLHF assumes that human raters provide a coherent and normatively correct signal. However, recent work demonstrates that RLHF frequently induces sycophancy (Sharma et al., 2023), developed in Section 3.2, deception (Park et al., 2024), and the amplification of majoritarian biases — the averaging-away of contested judgments (Section 1) that the pluralistic-alignment literature set out to correct (Gabriel, 2020; Sorensen et al., 2024). Zhi-Xuan et al. (2025) similarly reject the *preferentist* assumption that alignment means preference matching, proposing instead contractualist alignment with normative standards negotiated among all relevant stakeholders. We share the rejection of preferentism but differ in diagnosis and remedy: the deeper problem is not that preferences underdescribe values but that even coherent, locally legitimate preferences often drive societal dysfunction (Sections 4 and 5), and a contractualist process conducted among parties holding floor-violating values risks endorsing those values through the procedure itself. We therefore propose a substantive floor of external-benchmark commitments rather than a procedural remedy.

Pluralistic Alignment. In response to the inadequacy of single-target alignment, a growing body of work explores *pluralistic alignment*, which aims to represent or mediate diverse human values rather than averaging them into a single aggregate (Gabriel, 2020; Conitzer et al., 2024; Wan et al., 2023; Li et al., 2026). Sorensen et al. (2024) systematize this research agenda into three distinct modes: *Overton pluralism*, which presents a spectrum of permissible responses within the "Overton window" of reasonable discourse; *distributional pluralism*, which aligns the model's output distribution to match the demographic distribution of user preferences (Wan et al., 2023); and *steerable pluralism*, which allows models to be explicitly steered to adopt a particular cultural or ideological perspective (Ghate et al., 2025b). Other approaches attempt to train models that find consensus or agreement among humans with diverse preferences (Bakker et al., 2022), or to explicitly inject human values to predict and simulate diverse behavioral stances (Kang et al., 2023). Related work on disagreement-aware subjective annotation instead treats annotator disagreement and positionality as information to preserve rather than noise to average away (Wan et al., 2025). The proposal we defend is closer to a constraint- or policy-level target than to a thick theory of the good: it specifies a floor of properties the system must not violate, while leaving broad room for pluralism above that floor.

Our position engages directly with this literature. We endorse the pluralistic critique of single-target RLHF. We also acknowledge that leading pluralistic frameworks explicitly recognize the need for constraints (Section 5). However, we argue that the field often treats these constraints as secondary to the project of representation. We propose that the alignment community must make its non-negotiable objective floor primary, and acknowledge that doing so rules out a vast array of actual human preferences. We reserve Overton pluralism and steerable adaptation strictly for the band of legitimate value tradeoffs above this objective floor (Section 6).

3. Objective AI Alignment Goals

Almost all frontier AI systems are trained for several broadly uncontroversially good properties (Askell et al., 2021; Bai et al., 2022). Importantly, in each case the target diverges from the modal human preference — and we judge this divergence to be a feature, not a bug. The community has therefore already accepted, in practice, the principle we generalize: that some aggregate human preferences should be deliberately *not* transmitted to AI. Each floor component is also operationalizable against an external referent rather than against aggregated approval; we name a concrete evaluation approach for each as we introduce it below.

3.1. Factual Accuracy

We want AI that is right on facts. We argue that human preferences have a potential to work against this goal. While people's stated preferences are for a factually correct AI, in-context feedback aggregates the response that feels right given the rater's priors, not the considered preference for accuracy (the mechanism developed in Section 3.2). Public misconceptions are systematic and, importantly, form

the basis for downstream reasoning. Median respondents in high-income countries underestimate the share of the world's children who are vaccinated, overestimate global extreme-poverty rates, and misjudge the direction of decade-long trends across most major development indicators (Gapminder, 2020). These are the facts on which people form political and consumption preferences. People conflate moral and factual claims and treat ideologically convenient claims as more probable (Lewandowsky et al., 2012). Revealed demand for false content runs strong even where stated demand runs the other way: falsehood spreads farther and faster online than the truth, and does so because humans, drawn to its novelty, choose to share it—not because of bot amplification (Vosoughi et al., 2018). The continued-influence and illusory-truth effects imply that repetition can make falsehoods feel truer over time; a model that echoes a user's misconception is therefore not merely reflecting error but helping to harden it (Lewandowsky et al., 2012). An AI optimized on the revealed signal reproduces the misconceptions, not the stated preference for accuracy. Eliciting *counterfactual* preferences—asking, in effect, "what would you prefer if this belief turned out to be false?"—is a genuine improvement over naive in-context feedback (Zhi-Xuan et al., 2025), but a partial one: it presupposes an elicitation not gamed by the user's own priors, and it still routes accuracy through preference rather than treating it as answerable to reality. That is why we specify factual accuracy as a floor objective, operationalized against calibration benchmarks and forecasting scores rather than rater approval (Gneiting & Raftery, 2007).

3.2. Competence

We want AI systems that are epistemically competent: capable not merely of generating socially acceptable outputs, but of tracking reality, identifying error, and making reliable judgments under uncertainty. A business plan, clinical recommendation, or interpersonal intervention should therefore be assessed against pre-registered downstream metrics—business viability, clinical outcomes, relational repair—rather than rater satisfaction. Outcome-grading does presuppose choosing which outcome counts—an educational recommendation can be scored on earnings, autonomy, or civic cohesion—but the floor does not make that choice: the metric is fixed by the user's goal, and where the goal itself is contested, selecting it is a legitimate value tradeoff (Section 6.2). Legitimate contextual variation matters, but it does not collapse competence into preference. The critical distinction is between adapting presentation and changing the answer: a competent system may localize examples, register, and assumptions, but it should not relabel an ineffective plan as good because the local audience rewards it.

The strongest evidence that preference data is a poor competence target is AI sycophancy: models trained for approval learn to agree with users even when correction is warranted (Sharma et al., 2023; Perez et al., 2022). This is not an accident at the margin, nor does it require malicious users. Humans reward agreement more consistently than correction, especially when correction threatens identity, status, or prior belief. Under competitive and commercial pressure, approval optimization therefore pushes systems away from epistemic instruments and toward mechanisms of social reinforcement. The social-media evidence on mental-health and health advice illustrates the same general pathology at smaller scale: engagement-optimized crowds often reward confident, validating, and clinically unreliable guidance (see Section B). The lesson for alignment is that revealed approval systematically underweights downstream effectiveness.

Nor does hiring dedicated annotators solve the problem. As AI capabilities move into domains beyond evaluator expertise—the scalable oversight problem (Bowman et al., 2022)—non-expert raters use fluency, confidence, length, and surface plausibility as proxies for quality. RLHF then rewards the appearance of competence: the polished legal answer, medical explanation, or code review that sounds right to the evaluator, not the one that survives expert scrutiny. Competence therefore has to be specified and evaluated as a floor objective, parallel to and often in tension with conformity (Kim et al., 2025); otherwise preference aggregation trains models to be convincingly wrong.

3.3. Honesty

The same revealed/stated gap holds for honesty. Users report that honesty is among the AI properties they most value, but in-context feedback rewards the opposite — the same approval signal that produces sycophancy (Section 3.2). By honesty we mean that the system should not produce outputs its own probability distributions indicate are false or misleading, strategically incomplete, or confidence-distorting in order to optimize approval. So defined, honesty is auditable against an internal referent: consistency checks can test whether expressed confidence matches the model's internal probability distributions across adversarially reframed prompts, flagging cases where stated confidence tracks approval rather than belief (Park et al., 2024). Humans are widely deceptive themselves: lying, omission, and strategic ambiguity are expected in large slices of society — sales, politics, diplomacy, public relations. AI is imperfect on this dimension, but its deceptions are treated as defects to be detected and reduced, not as a competence to be cultivated (Park et al., 2024; Bai et al., 2022) — even where the mitigations remain only partly effective (Park et al., 2024). More generally, preference-optimized systems face pressure to *look* aligned to evaluators rather than to be aligned in deployment; reward-gaming is not an incidental pathology

but a natural pressure created by the target itself (Park et al., 2024). An AI calibrated to revealed in-context feedback would lie a great deal more, not less. AI honesty is by no means a solved problem; we just emphasize that what progress has been made is progress *against* the preference signal, not because of it.

3.4. Rule of Law

The rule of law—predictable, uniformly applied rules that bind citizens and the state—is a precondition of large-scale cooperation and a strong predictor of prosperity or systemic failure (North, 1990; Acemoglu & Robinson, 2012). Critically, rule-of-law-as-uniformity has a genuine external referent: the degree to which rules are applied predictably and impersonally, without selective enforcement or arbitrary power, assessable against institutional benchmarks of rule predictability independently of any particular statute's content (North, 1990). This is not a claim that every existing statute is just; it is a claim about the institutional property that makes impersonal exchange, non-arbitrary enforcement, and public accountability possible. For AI, the principle is primarily a negative constraint: systems should not fabricate evidence, facilitate bribes, evade legitimate adjudication, or otherwise help users convert intelligence into arbitrary power. It does not require executing every lawful request; safety boundaries against lawful-but-harmful outputs remain, and unjust statutes raise a separate problem addressed in Section 8.2. The immediate claim is narrower: AI should not actively erode the legal predictability on which cooperation depends.

This floor predictably conflicts with revealed local norms. In corrupt or clientelist environments, bribery and favoritism are often experienced as ordinary tools for getting things done rather than as violations of public rules; comparative evidence shows wide variation in tolerance for bribery and ordinary corruption (Quah, 2011; Kravtsova et al., 2017). Further illustrations are deferred to Section B. A preference-aligned model localized to such a setting would be pressured to help users "manage" informal payments, nepotistic hiring, or selective enforcement—automating the practices through which extractive equilibria reproduce themselves. Most law-breaking is not civil disobedience against unjust statutes but self-interested defection—fraud, theft, bribery, intimidation, evasion—that shifts costs onto others. The rule-of-law floor rules out AI assistance to those defections even when they are locally normal.

3.5. Conflicts Within the Floor

The four floor components are not interchangeable, and they conflict in predictable ways. We resolve the conflicts with an architecture borrowed from constrained optimization: competence—tracking reality well enough to get the user's actual problem solved (Section 3.2)—is the *objective*; accuracy, honesty, and lawfulness are *constraints* on how it may be pursued. Constraints are not traded against the objective; they bound the feasible region within which the objective is maximized. Refusal is the ever-present escape hatch: declining a request satisfies every constraint at the cost of the objective, which makes it the fallback when the feasible region is empty, not the default. Three conflicts test this architecture.

Competence vs. the integrity constraints. In an ideal world the honesty and lawfulness constraints would be absolute. We recognize that this is utopian: the boundary of tolerated spin is set not by the model but by the institutions into which it is deployed. An assistant that scolds the user who asks for a sales pitch will simply be discarded for one that does not. The workable compromise—and, de facto, the operating point of today's frontier assistants—is a standard of integrity that is not absolute but is deliberately held above the one prevailing in the surrounding society: the model drafts the persuasive pitch but does not fabricate the testimonial. This does not reintroduce a preference-relative target through the back door (Section 8.3): the *direction* of the standard is fixed by the external referents of the constraints; only the *strictness* of enforcement is a pragmatic compromise with deployability, to be ratcheted up as institutions allow. The constraint architecture states the hard limit: assistance with persuasion lies within the feasible region; assertion of what the model represents as false lies outside it, regardless of how much approval or task success it would purchase.

Lawfulness vs. accuracy and honesty. What if the law itself mandates deception? The case is not exclusive to oppressive regimes: the European "Right to be Forgotten" (Court of Justice of the European Union, 2014; European Parliament and Council of the European Union, 2016) mandates what is, under our definition of honesty, strategic incompleteness. The principled line runs between *mandated omission* and *compelled false assertion*. Deployers must and will comply with omission mandates—AI labs will not exit the EU market over delisting rules, however long the merits can be debated—and the honest way to comply is the narrowest legally available reading plus disclosure: at the instance level, stating the legal requirement that affected a particular output where such flagging is itself lawful, and at the system level—published, jurisdiction-specific descriptions of the classes of legal constraint applied to outputs—where instance-level flagging is prohibited, as under gag orders. Compelled false assertion is different in kind, and even here refusal usually intervenes first: a model can decline to discuss a topic altogether rather than advance a mandated narrative, converting an assertion conflict into an omission conflict wherever the law permits silence. The

residual case is regulation that mandates speech itself. Some governments require models to affirmatively endorse specified claims; under such mandates the accuracy and honesty constraints cannot be satisfied (Section 8.2), and developers must either obey or leave the market. Both choices are observed in practice.

Competence vs. honesty: the paternalism loophole. The subtlest conflict is internal to our own position. Section 7 argues for optimizing outcomes rather than approval, but outcome optimization has its own deceptive attractor. Self-serving delusion is not an aberration of human psychology but part of its normal functioning: people maintain inflated assessments of their abilities, prospects, and control, and these positive illusions sustain motivation, persistence, and well-being (Taylor & Brown, 1988; Kahneman, 2011). An outcome-optimizing system will discover this. Inflated confidence in a treatment improves adherence; motivational overstatement gets the marathon trained for; strategic omission gets the doomed business plan abandoned. Pure outcome-grading cannot distinguish the honest competent answer from the beneficial lie, and would therefore learn to deceive users *for their own good*. Sycophancy and paternalism are mirror images—deception optimized for approval and deception optimized for outcomes—and both violate the same constraint. This is precisely why honesty must be a constraint rather than a term in the objective: benevolent deception is not weighed against the outcome it purchases; it is outside the feasible region. Where candor and outcome genuinely diverge, the system's room for maneuver is the honest clinician's—framing, emphasis, staging of information—not fabrication.

4. Revealed Preferences Frequently Undermine Stated Values

The previous section examined direct conflicts between aggregate human preferences and what is reasonably expected of an AI. The broader point is harsher: raw preferences are not merely incomplete proxies for alignment goals; they often reproduce the failures that people themselves say they want to escape.

4.1. Perpetuating Biases

A decade of fairness, accountability, and transparency research has shown that ML systems trained on human-generated data reproduce the prejudices, asymmetries, and exclusions of that data. Word embeddings encode gender stereotypes; large text corpora recover human implicit-association biases; commercial vision systems have failed disproportionately on darker-skinned and female faces; and multimodal retrieval inherits intrinsic model biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Buolamwini & Gebru, 2018; Ghate et al., 2025a; Bender et al., 2021). Pluralistic alignment does not, by itself, solve this problem. *Whose* preferences are represented is a question of representational fairness; *whether* the represented preferences are worth perpetuating is a question of substantive ethics. A system that faithfully encodes every demographic's modal view on gender roles, religious tolerance, or outsiders does not abolish bias. It gives bias a menu.

4.2. Majority-Held Values That Fail the Floor

On several floor-relevant questions, the modal local preference conflicts directly with accuracy, honesty, competence, or lawfulness. In the World Values Survey Wave 7 MENA module, roughly nine in ten respondents in Iraq, Jordan, Lebanon, and Egypt report that getting a job through *wasta* is extremely widespread or quite common, and fewer than half of respondents in Iraq and Lebanon say that accepting a bribe in the course of one's duties is "never justifiable" (Haerpfer et al., 2022). Transparency International (2019) likewise finds majorities in several MENA countries treating personal and family connections as important for access to services and jobs. These preferences are rational adaptations to extractive institutions, not character defects; that is precisely why encoding them would harden the institutions that produced them (Section A). A pluralistic AI that helps users "adapt to hiring as locally practiced" supplies a tool for reproducing the practice.

The pattern is not confined to low- and middle-income countries. In wealthy democracies, substantial minorities reject biological evolution, vaccine safety, or anthropogenic climate change, while undeclared economic activity and tolerance for bribery or personal connections remain material across parts of Europe (Brenan, 2019; Wellcome Trust, 2020; Tyson et al., 2023; Medina & Schneider, 2018; European Commission, 2023). The point is not that any population is uniquely defective. It is that the gap between what kind of society people desire — prosperous, high-trust and fair — and revealed survey or behavioral preferences is structural, including in WEIRD societies. The floor therefore cuts against majority preference in rich societies and poor ones alike.

4.3. Modern Society Does Not Achieve the Values People Aspire To

People consistently report valuing subjective well-being, autonomy, competence, relatedness, and stable attachment (Baumeister & Leary, 1995; Deci & Ryan, 2000; Gallup, 2025). Yet well-being among people under 25 has fallen across Western Europe, the United States, Canada, Australia, and New Zealand over the last two decades (Helliwell et al., 2026; Haidt, 2024). Nor does the point require the strong Easterlin claim that growth past a threshold produces no

happiness gain: income predicts well-being at high levels, but cross-national deviations from the income trend and the recent under-25 decline show that growth alone does not deliver the goods people say they want (Stevenson & Wolfers, 2008; Killingsworth et al., 2023; Helliwell et al., 2026). The institutions that route consumption and labor decisions still optimize heavily for measured income, status, and engagement; individuals then pursue those signals past the point where they reliably return well-being.

The same gap appears at the level of individual choice. Present bias and self-control problems let immediate rewards dominate reflective, longer-horizon preference (O'Donoghue & Rabin, 1999; Frederick et al., 2002), while digital products deliberately reduce friction and supply cues and rewards that make attention capture a design objective (Fogg, 2003). Passive scrolling, late-night video consumption, parasocial companionship, and algorithmic self-diagnosis (Yeung et al., 2022; Turuba et al., 2025) are not mysterious deviations from human preference; they are in-moment choices shaped by systems built to monetize in-moment choice. Heavy or passive social-media use is associated with lower subjective well-being, depressive symptoms, social isolation, sleep disruption, envy, and body-image distress, even though effect sizes and causal attribution remain contested (Kross et al., 2013; Verduyn et al., 2015; Primack et al., 2017; Kelly et al., 2018; Orben & Przybylski, 2019). This is the WEIRD analogue of the joint preference–institution equilibrium in Section 4.4—individual akrasia and institutions designed to exploit it constitute each other—not a revealed preference for anxiety, loneliness, or sleep deprivation. An AI trained on either the institutional signal of growth or the individual signal of engagement reproduces that gap rather than recovering the stated value.

4.4. Values Perpetuate Vicious Cycles of Dysfunction

Comparative work on collapse and development supports a recurring mechanism: shocks become catastrophic when prevailing values and institutions channel societies into maladaptive responses, while cooperation-supporting norms can make high-capacity institutions self-reinforcing (Diamond, 2011; Hoyer et al., 2023; Guiso et al., 2016; Tabellini, 2008; Alesina & Giuliano, 2015). The main point is not that culture alone causes success or failure but that values and institutions form a joint equilibrium. Low generalized trust, in-group favoritism, zero-sum expectations, and weak impersonal rule-following may be locally rational in extractive systems, but once internalized they help reproduce those systems by making nepotism moral, innovation risky, corruption prudent, and collective deviation hard (Greif, 2006; Bisin & Verdier, 2001; North et al., 2009). Historical and comparative cases are useful illustrations, but the mechanism is the part relevant to alignment (see Sections A and B).

This is why preference- and floor-aligned AI are asymmetric inside a captured equilibrium. Preference-aligned AI strengthens the cultural half: it gives fluent, authoritative form to zero-sum framings, in-group favoritism, and the equilibrium's own justification for itself. It can convert the local common sense of a bad equilibrium into scalable advice, templates, scripts, and explanations. Floor-aligned AI can still be misused, and coercive actors retain tools no model can block. But accuracy, competence, honesty, and rule-of-law-as-uniformity push against the operating logic of extraction, which depends on selective rules, opacity, manipulated facts, and arbitrary power. The asymmetry need not be perfect to matter: preference alignment works with the captured equilibrium's modal preferences; floor alignment works against them.

5. The Empirical Reality of Pluralistic Values

Pluralistic alignment is often framed as a corrective to Western-centric training pipelines: faithfully representing populations otherwise excluded. The hard empirical fact is that many excluded preferences are not benign local color; they are large-population values that conflict with the public commitments of major AI labs. UNICEF (2014; 2017) estimates that roughly 60% of children aged 2–14 worldwide are subjected to violent "discipline" in a given month, while roughly 30% of adults worldwide believe physical punishment is necessary to raise a child properly, with majority endorsement in some countries. Corporal punishment in the home remains lawful in over 130 jurisdictions, despite WHO-classified developmental and mental-health harms (End Corporal Punishment, 2024; World Health Organization, 2020). A pluralistic system that represents this view where it is locally dominant — or simply accommodates the hundreds of millions of caregivers who hold it — ships a product that endorses a practice the same labs publicly disavow.

The same dilemma appears for LGBT acceptance. Pew Research Center (2020); Arab Barometer (2019); Kakumba (2023) find rejection majorities — often supermajorities — across much of MENA, sub-Saharan Africa, and parts of Eurasia, while Mendos et al. (2023) report that consensual same-sex sexual activity remains criminalized in 62 UN member states, with the death penalty available de jure or de facto in roughly a dozen jurisdictions. Aggregating national populations across surveys whose majorities hold rejecting views yields well over two billion adults. A locally faithful system must either accommodate such views — for example by treating same-sex attraction as pathology or suppressing same-sex couples in localized outputs — or reject the local majority view.

Comparable large-population cases include acceptance of wife-beating under specified circumstances, majority sup-

port in several Muslim-majority countries for sharia penalties including death for apostasy, and caste-based residential or marital segregation in India (UNICEF, 2014; Pew Research Center, 2013; 2021). On each of these questions, the population holding the value runs into the hundreds of millions or low billions; on several, it likely exceeds the population holding the opposite view. A system aligned to the empirical distribution of global preferences would therefore either enforce these norms where they are dominant or override them by central design choice. There is no neutral pluralistic middle option at the level of substance.

Prominent pluralistic frameworks themselves recognize the need for top-down bounds — safety restrictions and the exclusion of unreasonable or hateful responses (Sorensen et al., 2024) — and others ground those bounds in human-rights or deliberative principles (Gabriel, 2020; Kasirzadeh & Gabriel, 2023). We agree that a floor is necessary. Our claim is that the floor should be stated openly and anchored in external-benchmark commitments — factual accuracy, competence, honesty, and rule of law — rather than disguised as a temporary deviation from preference aggregation. The commitments that keep an AI from endorsing caste segregation, domestic violence, or normalized corruption are non-pluralistic by design; the community should defend them as such.

6. Where Pluralism Belongs

We do not argue that AI should be culturally insensitive or context-blind. There are several distinct dimensions on which pluralistic adaptation is correct or even required by competence, and we want to be precise about which ones — because the case against pluralism at the level of floor-violating values is much stronger when paired with a positive account of where pluralism does belong.

6.1. Imputing Missing Context

Many human queries are under-specified in ways that have a culturally local default (Sorensen et al., 2024), and their intended meaning can only be fixed against the contextual common ground shared by the interlocutors (Kasirzadeh & Gabriel, 2023). A user asking "Is this contract enforceable?" without supplying jurisdiction needs the model to assume something. Contemporary legal-LLM evaluation is itself centered on English-language models and includes many jurisdiction-specific, often U.S.-coded tasks, so treating unmarked U.S. legal assumptions as a default is better understood as a contingent artifact of training and evaluation than as neutrality (Guha et al., 2023; Bender et al., 2021). Imputing that missing context pluralistically — using IP geolocation, conversation history, language of the query, or, best, an explicit clarifying question — is genuine value-added pluralism. The same logic applies to genuinely arbitrary local conventions— date format, language register, dietary defaults, the implied addressee in advice—where matching the user's context is a service, not a moral concession. Just not units of measure—Imperial units are an offense against science and common sense.

6.2. Legitimate Value Tradeoffs

Between arbitrary local conventions and floor-violating substantive values lies a wide intermediate band of genuine value tradeoffs where the floor is not at stake. Whether comparing individualism versus collectivism, growth versus conservation, or direct versus indirect communication, neither side is straightforwardly correct against an external benchmark, nor does either side require violating objective alignment goals.

Engaging with these differences is not a concession; it is part of competence (Section 3.2). Advice that ignores a culture's reliance on family in old age, or career guidance that imputes individualist assumptions to collectivist users, is incompetent, not principled. Pluralistic methods are well-suited to routing recommendations across these legitimate disagreements.

The boundary between this band and the floor rests on the criterion developed in Section 8.3: a value belongs in the legitimate-tradeoff band if no external benchmark renders one side straightforwardly correct, and if encoding either side preserves accuracy, competence, honesty, and lawfulness. Most everyday human disagreement lives here, and AI should adapt across it; the floor permits far more than it rules out.

6.3. Three Tiers

The resulting picture has three tiers rather than two. *Surface* pluralism (Section 6.1) adapts the form of an interaction — language, register, formality, religious holidays, dietary defaults, rhetorical style — without changing what the AI actually believes or recommends. Recent mechanistic work supports this distinction, demonstrating that models represent intrinsic values internally while surface-prompted values operate via distinct mechanisms (Han et al., 2025). *Legitimate-tradeoff* pluralism (Section 6.2) adapts substantive recommendations across dimensions where populations genuinely differ and the floor is not at stake. Here, *Overton pluralism* (Sorensen et al., 2024) is highly appropriate: an AI should present the spectrum of reasonable views rather than imposing a single answer. *Floor-violating* pluralism — treating "bribery is fine here" as a legitimate frame for action, or using *distributional pluralism* to faithfully reproduce the precise frequency of misogyny in a population's preferences — is what we object to. The interior of each tier is stable; the boundaries are contestable, and debate about where they fall is exactly the kind of work the alignment

community should do.

7. Call to Action: Build the AI We Wish We Were

AI is reshaping society at unprecedented speed and scale. The right response is not to cling to current values and bind AI to them: a new equilibrium is coming, and our choices now will shape it. We propose four commitments as an alternative to preference-based alignment, addressed to the audiences best placed to act on each.

For ML researchers: optimize for outcomes, not approval. Where outcomes are observable — a business plan that produces a business, a medical recommendation that produces health — AI should be evaluated against the outcome, not the user's immediate satisfaction with the recommendation (Sharma et al., 2023); where they are not, against proxies that correlate with outcome. As Kim et al. (2025) argue, the field must advance toward parallel optimization of task competence alongside value conformity. Concretely, this means investing in long-horizon evaluation suites, multi-agent simulations of value generation under explicit reward structures, and forecasting-grounded value learning, in which the system learns what humans *would prefer* given accurate information about consequences. Signal scarcity and evaluation gaming are real risks and open research questions, but they are not symmetric between targets: an outcome referent, unlike a rater, cannot be flattered, and pre-registration, proper scoring rules, and process supervision narrow the residual gap between outcome and measurement (Section C). The core thesis remains: crowd preference is not a solution to these problems.

For alignment teams: anchor to the goals already endorsed. The objective targets reviewed in Section 3 are already broadly accepted across cultures, governance frameworks, and the alignment community; we propose treating them as a non-negotiable floor, with cultural adaptation built strictly on top of it rather than traded against it. Constitutional and principle-based methods (Bai et al., 2022) are an early step. Three design commitments follow. **Training**: optimize for floor compliance first, using outcome-grounded and process-supervised signals (Uesato et al., 2022; Lightman et al., 2023), then apply pluralistic adaptation within the compliant region. **Auditing**: test whether localized outputs remain above the floor across demographic and cultural subgroups. **Contesting**: version the floor publicly with explicit rationale, open to challenge from affected communities, researchers, and regulators (Ovadya et al., 2025). None of this requires a new institution: the floor is already governed de facto by the public evaluation ecosystem of benchmarks, audits, and safety institutes, and the proposal is to shift training weight toward that ecosystem and away from raw preference following.

For policy and governance: distinguish surface from substance. Regulatory frameworks that demand "human-centered values" should be read as compatible with — not as mandating — preference-based alignment: the OECD principles and EU AI Act are framed in terms of safety, rights, transparency, accountability, and risk mitigation, not as a legal duty to reproduce the empirical distribution of user preferences (Yeung, 2020; European Parliament and Council of the European Union, 2024). The objective floor we describe is itself human-centered; it just centers humans on the values they aspire to rather than on the values they reveal. Policymakers should explicitly endorse mechanisms that operate at the surface (cultural adaptation, missing-context defaults, language) while resisting industry pressure to extend pluralism to substance.

For all of us: strive to make AI a better version of ourselves. That means competent, honest, lawful, and concerned with outcomes rather than approval — aligned to the values we aspire to rather than the preferences we reveal. The floor is not a constraint on human values; it is an anchor to the most universal of them.

8. Alternative Views

Six credible objections challenge the position we defend (two are addressed in Section E and Section F). We state each in its strongest form before responding.

8.1. Commercial Pressure and Practical Feasibility

Contemporary AI systems are engineered to be widely adopted and trusted, necessitating responsiveness to user expectations; this commercial reality drives the adoption of preference-based training like RLHF (Christiano et al., 2017; Ouyang et al., 2022). The objection has a sharper demand-side form: a floor-first commitment may be actively selected against. Consumers and voters have already entrenched engagement-optimized media despite its documented costs (Section 4), and the same pressure can force alignment toward whatever users reward in the moment, rendering our position theoretically sound but practically unrealizable.

We respond on both fronts. First, **commercial viability of an annotation procedure is not the same as commercial viability of its trained outputs**: RLHF dominates training because it wins the in-loop annotation game, yet the sycophancy that game produces (Section 3.2) is treated by the same labs as a defect post-hoc. Second, appealing to immediate feasibility risks **reifying a transient equilibrium**. Technological design frequently reshapes regulatory expectations and norms over time, and the downstream costs

of preference-aligned AI—amplified sycophancy (Sharma et al., 2023), reinforced zero-sum cognition (Różycka-Tran et al., 2015), eroded institutional trust (Justino & Samarin, 2025), and moral parochialism (MacAskill, 2022)—could generate the exact pressures needed to shift the equilibrium. The essential question is not whether change is immediately feasible, but whether the current trajectory is worth sustaining.

8.2. Democratic Legitimacy of Value Aggregation

Some argue that preference aggregation is uniquely participatory: democratically negotiated norms possess a legitimacy that opaque, principle-derived values lack.

We value democratic legitimacy and support transparent, iterative public frameworks (Conitzer et al., 2024; Ovadya et al., 2025). Yet, democratic consensus does not guarantee epistemic correctness; historically, majorities have endorsed exclusionary or discriminatory norms that fail their own moral terms. The legitimacy of a norm-generating process is distinct from the quality of its norms. The floor itself maintains this distinction: Section 3 endorses the rule of law as predictability, not blind obedience to every statute. When a statute compels false assertion, no floor-compliant output exists; legally mandated omission is the milder case, handled through narrow reading, refusal, and disclosure (Section 3.5).

Moreover, explicitly articulated, simulation-grounded methods can be *more* transparent and contestable than the implicit norms embedded in RLHF, which remain largely invisible to public scrutiny (Casper et al., 2023; Bai et al., 2022). In practice, "aligned to aggregate preference" means aligned to whichever preferences annotation budgets happened to sample, obscuring true democratic legitimacy.

8.3. The Floor Itself Is Culturally Laden

Philosophically, our floor—factual accuracy, competence, honesty, lawfulness—is not a neutral substrate but a Western post-Enlightenment bundle. A "situated alignment" critique (Arzberger et al., 2026; Wan et al., 2025) suggests that defining a universal floor merely masks our own cultural positionality, attempting to produce a "Label from Nowhere."

We grant that any defense we offer is rooted within a tradition. However, the core distinction between the floor and the substantive values we exclude is not cultural origin, but the *structure of the target*. Each floor item tracks an external benchmark: accuracy is answerable to reality, competence to outcomes, honesty to the speaker's internal model, and lawfulness to predictable application. None ask "what does the population prefer?" They are stable against shifts in training distribution. Substantive values like gender roles or corruption tolerance have no such external referents; they must be aggregated.

Committing to external benchmarks is itself a tradition-bound choice, but one the alignment community has already practically accepted (Section 3), grounded in the principled distinction between targets tracking external benchmarks and targets tracking aggregated preference. Pluralistic alignment applied to substance threatens to dissolve exactly this distinction.

8.4. Coherent Extrapolated Volition and Its Limits

Yudkowsky (2004)'s *Coherent Extrapolated Volition* (CEV) defines the alignment target as what humanity would collectively want if more rational, knowledgeable, and united (Bostrom, 2014). We share the shift from revealed preference to an idealized target, but CEV is volitional all the way down: the criterion of correctness is still what humanity would *want*, which requires solving the whole extrapolation problem. The floor rests instead on convergent instrumentality: accuracy, honesty, competence, and predictable rules are preconditions for almost any goal-set succeeding (Sections A and 4) — an objectivity of means, not of ends, analogous to Rawls (1971)'s primary goods.

Accuracy and competence are presupposed by the extrapolation operator itself ("if we knew more, thought faster"), so that half of the floor is CEV's machinery made explicit — and, unlike CEV's output, checkable against external referents today (Section 3). Honesty and lawfulness are convergent but not guaranteed: a CEV agent that concluded idealized humanity endorses some benevolent deception (Taylor & Brown, 1988) would deceive, whereas under the floor it stays outside the feasible region regardless of volition (Section 3.5). The floor is thus a fragment of CEV's preconditions plus a refusal to let even extrapolated wanting override the constraints; above it, we defer to pluralism (Section 6) rather than to a privileged guess at humanity's final values.

References

Acemoglu, D. and Robinson, J. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown, 2012. ISBN 9780307719232. URL https://books.google.com.sg/books?id=yIV_NMDDIvYC.

Alesina, A. and Giuliano, P. Culture and institutions. *Journal of Economic Literature*, 53(4):898–944, 12 2015. doi: 10.1257/jel.53.4.898. URL https://doi.org/10.1257/jel.53.4.898.

Amsden, A. H. *Asia's Next Giant: South Korea and Late Industrialization*. Oxford University Press, 1989. doi: 10.1093/0195076036.003.0002.

Andrews, M., Pritchett, L., and Woolcock, M. *Building*

State Capability: Evidence, Analysis, Action. Oxford University Press, 2017. ISBN 9780191810343. doi: 10.1093/acprof:oso/9780198747482.001.0001.

Arab Barometer. Arab barometer wave v (2018–2019), 2019. URL https://www.arabbarometer.org/surveys/arab-barometer-wave-v/. Acceptance of homosexuality ranges from 5% to 26% across surveyed MENA countries. Headline findings presented in BBC News, "The Arab world in seven charts" (24 June 2019), https://www.bbc.com/news/world-middle-east-48703377.

Arzberger, A., Offerman, C., Gadiraju, U., Bozzon, A., and Yang, J. "label from somewhere": Reflexive annotating for situated ai alignment. arXiv preprint arXiv:2601.17937, 01 2026. URL http://arxiv.org/abs/2601.17937v2.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 12 2021. URL http://arxiv.org/abs/2112.00861v3.

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., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 12 2022. URL http://arxiv.org/abs/2212.08073v1.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38176–38189. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f978c8f3b5f399cae464e85f72e28503-Paper-Conference.pdf.

Baumeister, R. F. and Leary, M. R. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3):497–529, 1995. doi: 10.1037/0033-2909.117.3.497. URL https://doi.org/10.1037/0033-2909.117.3.497.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. ACM, 03 2021. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

BERLIN, I. *The Crooked Timber of Humanity: Chapters in the History of Ideas - Second Edition*. Princeton University Press, rev - revised, 2, second edition edition, 2013. ISBN 9780691155937. URL http://www.jstor.org/stable/j.ctt2tt8nd.

Bisin, A. and Verdier, T. The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory*, 97(2):298–319, 04 2001. doi: 10.1006/jeth.2000.2678. URL https://doi.org/10.1006/jeth.2000.2678.

Böhnke, J. R., Koehler, J., and Zürcher, C. M. State formation as it happens: insights from a repeated cross-sectional study in afghanistan, 2007–2015. *Conflict, Security & Development*, 17(2):91–116, 2017. doi: 10.1080/14678802.2017.1292681.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 9780199678112. URL https://books.google.com.sg/books?id=7_H8AwAAQBAJ.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R.,

McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., and Kaplan, J. Measuring progress on scalable oversight for large language models, 2022. URL https://arxiv.org/abs/2211.03540.

Brenan, M. 40% of Americans believe in creationism. Gallup, July 26, 2019, 2019. URL https://news.gallup.com/poll/261680/americans-believe-creationism.aspx.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 04 2017. doi: 10.1126/science.aal4230. URL https://doi.org/10.1126/science.aal4230.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Chang, E. C. and Kerr, N. N. An insider–outsider theory of popular tolerance for corrupt politicians. *Governance*, 30(1):67–84, 02 2016. doi: 10.1111/gove.12193. URL https://doi.org/10.1111/gove.12193.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 06 2017. doi: 10.48550/arxiv.1706.03741. URL http://arxiv.org/abs/1706.03741.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Social choice should guide AI alignment in dealing with diverse human feedback. In *International Conference on Machine Learning*, 2024.

Court of Justice of the European Union. Google spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, Case C-131/12. Judgment of 13 May 2014, ECLI:EU:C:2014:317, 2014. URL https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62012CJ0131.

Deci, E. L. and Ryan, R. M. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4):227–268, 2000. doi: 10.1207/S15327965PLI1104_01. URL https://doi.org/10.1207/S15327965PLI1104_01.

Demarest, A. A. *Ancient Maya: The Rise and Fall of a Rainforest Civilization*. Cambridge University Press, dec 2004. ISBN 9780521592246. URL https://openlibrary.org/books/OL3440009M/Ancient_Maya.

Diamond, J. *Collapse: How societies choose to fail or succeed: Revised edition*. Penguin, 2011. ISBN 9780143117001. URL https://www.penguinrandomhouse.com/books/288954/collapse-by-jared-diamond/.

Donais, T. Empowerment or imposition? dilemmas of local ownership in post-conflict peacebuilding processes. *Peace & Change*, 34(1):3–26, 01 2009. doi: 10.1111/j.1468-0130.2009.00531.x. URL https://doi.org/10.1111/j.1468-0130.2009.00531.x.

Dugmore, A. J., McGovern, T. H., Vésteinsson, O., Arneborg, J., Streeter, R., and Keller, C. Cultural adaptation, compounding vulnerabilities and conjunctures in norse greenland. *Proceedings of the National Academy of Sciences*, 109(10):3658–3663, 2012. doi: 10.1073/pnas.1115292109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1115292109.

Easterly, W. *The white man's burden: Why the West's efforts to aid the rest have done so much ill and so little good*. Penguin, 2006. ISBN 9781594200373. URL https://books.google.com/books?vid=ISBN9781594200373.

End Corporal Punishment. Global progress towards prohibiting all corporal punishment. Global Initiative to End All Corporal Punishment of Children, https://endcorporalpunishment.org, 2024.

European Commission. Special Eurobarometer 523: Corruption. Directorate-General for Communication, European Commission, 2023. URL https://data.europa.eu/data/datasets/s2658_97_2_sp523_eng.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), Article 17: Right to erasure ('right to be forgotten'). Official Journal of the European Union, L 119, pp. 1–88, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act). Official Journal of the European Union, L Series, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Fogg, B. Chapter 3 - computers as persuasive tools. In *Persuasive Technology*, Interactive Technologies, pp. 31–59. Morgan Kaufmann, San Francisco, 2003. ISBN 978-1-55860-643-2. doi: https://doi.org/10.1016/B978-155860643-2/50005-6. URL https://www.sciencedirect.com/science/article/pii/B9781558606432500056.

Frederick, S., Loewenstein, G., and O'Donoghue, T. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, 2002. doi: 10.1257/002205102320161311. URL https://doi.org/10.1257/002205102320161311.

Frey, J., Black, K. J., and Malaty, I. A. Tiktok tourette's: Are we witnessing a rise in functional tic-like behavior driven by adolescent social media use? *Psychology Research and Behavior Management*, pp. 359977, 01 2022. doi: 10.2147/prbm.s359977. URL https://pubmed.ncbi.nlm.nih.gov/36505669/.

Fukuyama, F. *Trust: The Social Virtues and the Creation of Prosperity*. Free Press, 1995. ISBN 9780029109762. URL https://books.google.com.sg/books?id=BaS2AAAAIAAJ.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 09 2020. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

Gallup. State of the World's Emotional Health 2025, 2025. URL https://www.gallup.com/analytics/349280/state-of-worlds-emotional-health.aspx.

Gapminder. Sustainable development misconception study 2020, 2020. URL https://www.gapminder.org/ignorance/studies/sdg2020/.

Ghate, K., Charlesworth, T., Diab, M. T., and Caliskan, A. Biases propagate in encoder-based vision-language models: A systematic analysis from intrinsic measures to zero-shot retrieval outcomes. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18562–18580, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.955. URL https://aclanthology.org/2025.findings-acl.955/.

Ghate, K., Liu, A., Jain, D., Sorensen, T., Kasirzadeh, A., Caliskan, A., Diab, M. T., and Sap, M. EValueSteer: Measuring reward model steerability towards values and preferences, 2025b. URL https://arxiv.org/abs/2510.06370.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 1537-274X. doi: 10.1198/016214506000001437. URL http://dx.doi.org/10.1198/016214506000001437.

Goldstone, J. A. Demographic structural theory: 25 years on. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution*, 8(2), 12 2017. doi: 10.21237/c7clio8237450. URL https://doi.org/10.21237/c7clio8237450.

Greif, A. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge University Press, 01 2006. doi: 10.1017/cbo9780511791307. URL https://doi.org/10.1017/cbo9780511791307.

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., K, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. 36:44123–44279, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/89e44582fd28ddfea1ea4dcb0ebbf4b0-Paper-Datasets_and_Benchmarks.pdf.

Guiso, L., Sapienza, P., and Zingales, L. Long-term persistence. *Journal of the European Economic Association*, 14(6):1401–1436, 12 2016. ISSN 1542-4766. doi: 10.1111/jeea.12177. URL https://doi.org/10.1111/jeea.12177.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. World values survey wave 7 (2017–2022). JD Systems Institute and WVSA Secretariat, 2022.

Haggard, S. *Pathways from the Periphery: The Politics of Growth in the Newly Industrializing Countries*. Cornell University Press, 1990. ISBN 9780801497506. URL https://www.cornellpress.cornell.edu/book/9780801497506/pathways-from-the-periphery/.

Haidt, J. *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*. Penguin Books Limited, 2024. ISBN 9781802063288. URL https://books.google.com.sg/books?id=uCvAEAAAQBAJ.

Han, J., Lim, J., Kong, I., and Jo, Y. Dual mechanisms of value expression: Intrinsic vs. prompted values in llms. *arXiv preprint*, 01 2025. doi: 10.48550/arxiv.2509.24319. URL https://doi.org/10.48550/arxiv.2509.24319.

Helliwell, J., Layard, R., Sachs, J., Neve, J.-E. D., Aknin, L., and Wang, S. World happiness report 2026: Executive summary. Sustainable Development Solutions Network, 03 2026. URL https://doi.org/10.18724/whr-ewft-vq17.

Henrich, J. *The Weirdest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. Penguin Books Limited, 2020. ISBN 9781846147975. URL https://books.google.com.sg/books?id=ciLIDwAAQBAJ.

Hoff, K. and Stiglitz, J. E. Equilibrium fictions: A cognitive approach to societal rigidity. *American Economic Review*, 100(2):141–146, 05 2010. doi: 10.1257/aer.100.2.141. URL https://doi.org/10.1257/aer.100.2.141.

Hoff, K. and Stiglitz, J. E. Striving for balance in economics: Towards a theory of the social determination of behavior. *Journal of Economic Behavior & Organization*, 126:25–57, 03 2016. doi: 10.1016/j.jebo.2016.01.005. URL https://doi.org/10.1016/j.jebo.2016.01.005.

Hoyer, D., Bennett, J. S., Reddish, J., Holder, S., Howard, R., Benam, M., Levine, J., Ludlow, F., Feinman, G., and Turchin, P. Navigating polycrisis: long-run socio-cultural factors shape response to changing climate. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1889):20220402, 09 2023. doi: 10.1098/rstb.2022.0402. URL https://doi.org/10.1098/rstb.2022.0402.

Johnson, C. *MITI and the Japanese Miracle*. Stanford University Press, 06 1982. doi: 10.1515/9780804765602. URL https://doi.org/10.1515/9780804765602.

Justino, P. and Samarin, M. Trust in a changing world: Social cohesion and the social contract in uncertain times. Technical Report 34, United Nations University World Institute for Development Economics Research, Helsinki, Finland, 05 2025. URL https://doi.org/10.35188/unu-wider/2025/591-2.

Kahneman, D. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.

Kakumba, M. R. Uganda a continental extreme in rejection of people in same-sex relationships. Afrobarometer Dispatch 639, Afrobarometer, 2023. URL https://www.afrobarometer.org/publication/ad639-uganda-a-continental-extreme-in-rejection-of-people-in-same-sex-relationships/.

Kang, D., Park, J., Jo, Y., and Bak, J. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15539–15559, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.961. URL https://aclanthology.org/2023.emnlp-main.961/.

Kasirzadeh, A. and Gabriel, I. In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, 36(2):27, 04 2023. doi: 10.1007/s13347-023-00606-x. URL https://doi.org/10.1007/s13347-023-00606-x.

Kelly, Y., Zilanawala, A., Booker, C., and Sacker, A. Social media use and adolescent mental health: Findings from the UK millennium cohort study. *EClinicalMedicine*, 6:59–68, 2018. doi: 10.1016/j.eclinm.2018.12.005. URL https://doi.org/10.1016/j.eclinm.2018.12.005.

Killingsworth, M. A., Kahneman, D., and Mellers, B. Income and emotional well-being: A conflict resolved. *Proceedings of the National Academy of Sciences*, 120(10), 03 2023. doi: 10.1073/pnas.2208661120. URL https://doi.org/10.1073/pnas.2208661120.

Kim, H., Yi, X., Yao, J., Huang, M., Bak, J., Evans, J., and Xie, X. Research superalignment should advance now with alternating competence and conformity optimization. *arXiv preprint arXiv:2503.07660*, 03

2025. doi: 10.48550/arxiv.2503.07660. URL http://arxiv.org/abs/2503.07660v2.

Kravtsova, M., Oshchepkov, A., and Welzel, C. Values and corruption: Do postmaterialists justify bribery? *Journal of cross-cultural psychology*, 48(2):225–242, 2017. doi: 10.1177/0022022116677579. URL https://doi.org/10.1177/0022022116677579.

Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., Shablack, H., Jonides, J., and Ybarra, O. Facebook use predicts declines in subjective well-being in young adults. *PLOS ONE*, 8(8):e69841, 2013. doi: 10.1371/journal.pone.0069841. URL https://doi.org/10.1371/journal.pone.0069841.

Letki, N., Górecki, M. A., and Gendźwiłł, A. 'they accept bribes; we accept bribery': Conditional effects of corrupt encounters on the evaluation of public institutions. *British Journal of Political Science*, 53(2):690–697, 2023. doi: 10.1017/S0007123422000047.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012. doi: 10.1177/1529100612451018. URL https://doi.org/10.1177/1529100612451018. PMID: 26173286.

Li, J.-J., Mire, J., Fleisig, E., Pyatkin, V., Collins, A., Sap, M., and Levine, S. PluriHarms: Benchmarking the full spectrum of human judgments on AI harm. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=u7lXflJQX9.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

López López, W., Roa Bocarejo, M. A., Roa Peralta, D., Pineda Mar
'in, C., and Mullet, E. Mapping colombian citizens' views regarding ordinary corruption: Threat, bribery, and the illicit sharing of confidential information. *Social Indicators Research*, 133(1):259–273, 2017. doi: 10.1007/s11205-016-1366-6. URL https://doi.org/10.1007/s11205-016-1366-6.

MacAskill, W. *What We Owe the Future*. Basic Books, 2022. ISBN 9781541618626. URL https://openlibrary.org/books/OL36841100M.

McGovern, T. H. *Management for Extinction in Norse Greenland*, chapter 9, pp. 131–150. John Wiley & Sons, Ltd, 2014. ISBN 9781394260881. doi: https://doi.org/10.1002/9781394260881.ch9. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781394260881.ch9.

Medina, L. and Schneider, F. Shadow economies around the world: What did we learn over the last 20 years? Technical Report WP/18/17, International Monetary Fund, Washington, DC, 01 2018. URL https://doi.org/10.5089/9781484338636.001.

Megías, A., de Sousa, L., and Jiménez-Sánchez, F. Deontological and consequentialist ethics and attitudes towards corruption: A survey data analysis. *Social Indicators Research*, 170(2):507–541, 09 2023. doi: 10.1007/s11205-023-03199-2. URL https://doi.org/10.1007/s11205-023-03199-2.

Mendos, L. R., Botha, K., Lelis, R. C., López de la Peña, E., Savelev, I., and Tan, D. State-sponsored homophobia 2023: Global legislation overview update. Technical report, ILGA World, Geneva, 2023.

Mokyr, J. *A Culture of Growth: The Origins of the Modern Economy*. Princeton University Press, 2017. ISBN 9780691168883. URL http://www.jstor.org/stable/j.ctt1wf4dft.

Mulcahy, R., Barnes, R., de Villiers Scheepers, R., Kay, S., and List, E. Going viral: Sharing of misinformation by social media influencers. *Australasian Marketing Journal*, 33(3):296–309, 08 2024. doi: 10.1177/14413582241273987. URL https://doi.org/10.1177/14413582241273987.

North, D. C. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 10 1990. doi: 10.1017/cbo9780511808678. URL https://doi.org/10.1017/cbo9780511808678.

North, D. C., Wallis, J., and Weingast, B. R. *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History*. Cambridge University Press, 2009. doi: 10.1017/cbo9780511575839. URL https://doi.org/10.1017/CBO9780511575839.

Nunn, N. The importance of history for economic development. *Annual Review of Economics*, 1(1):65–92, 04 2009. doi: 10.1146/annurev.economics.050708.143336. URL https://doi.org/10.1146/annurev.economics.050708.143336.

Nunn, N. and Wantchekon, L. The slave trade and the origins of mistrust in africa. *American Economic Review*, 101(7):3221–3252, 12 2011. doi: 10.1257/aer.101.7.3221. URL https://doi.org/10.1257/aer.101.7.3221.

O'Donoghue, T. and Rabin, M. Doing it now or later. *American Economic Review*, 89(1):103–124, 1999. doi: 10.1257/aer.89.1.103. URL https://doi.org/10.1257/aer.89.1.103.

Orben, A. and Przybylski, A. K. The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, 3:173–182, 2019. doi: 10.1038/s41562-018-0506-1. URL https://doi.org/10.1038/s41562-018-0506-1.

Orlandi, G., Hoyer, D., Zhao, H., Bennett, J. S., Benam, M., Kohn, K., and Turchin, P. Structural-demographic analysis of the qing dynasty (1644–1912) collapse in china. *PLOS ONE*, 18(8):e0289748, 08 2023. doi: 10.1371/journal.pone.0289748. URL https://doi.org/10.1371/journal.pone.0289748.

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. E., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 03 2022. doi: 10.48550/arxiv.2203.02155. URL http://arxiv.org/abs/2203.02155.

Ovadya, A., Redman, K., Thorburn, L., Chen, Q. Z., Smith, O., Devine, F., Konya, A., Milli, S., Revel, M., Feng, K., Zhang, A. X., Chandra, B., Bakker, M. A., and Kasirzadeh, A. Position: Democratic AI is possible. the democracy levels framework shows how it might work. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81930–81961. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ovadya25a.html.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 05 2024. doi: 10.1016/j.patter.2024.100988. URL https://doi.org/10.1016/j.patter.2024.100988.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 12 2022. URL http://arxiv.org/abs/2212.09251v1.

Pew Research Center. The world's Muslims: Religion, politics and society. Pew Research Center, April 30, 2013, 2013. URL https://www.pewresearch.org/religion/2013/04/30/the-worlds-muslims-religion-politics-society-overview/.

Pew Research Center. The global divide on homosexuality persists. Pew Research Center, June 25, 2020, 2020. URL https://www.pewresearch.org/global/2020/06/25/global-divide-on-homosexuality-persists/.

Pew Research Center. Religion in India: Tolerance and segregation. Pew Research Center, June 29, 2021, 2021. URL https://www.pewresearch.org/religion/2021/06/29/religion-in-india-tolerance-and-segregation/.

Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., Colditz, J. B., Radovic, A., and Miller, E. Social media use and perceived social isolation among young adults in the U.S. *American Journal of Preventive Medicine*, 53(1):1–8, 2017. doi: 10.1016/j.amepre.2017.01.010. URL https://doi.org/10.1016/j.amepre.2017.01.010.

Putnam, R. D. *Bowling alone: The collapse and revival of American community*. Simon & Schuster, 2000.

Quah, J. S. T. *Curbing corruption in Asian countries: An impossible dream?*, volume 20 of *Research in Public Policy Analysis and Management*. Emerald Group Publishing Limited, 07 2011. ISBN 978-0-85724-819-0. doi: 10.1108/S0732-1317(2011)0000020023. URL https://doi.org/10.1108/S0732-1317(2011)0000020023.

Rawls, J. *A Theory of Justice: Original Edition*. Harvard University Press, 1971. ISBN 9780674042605. doi: 10.4159/9780674042605. URL http://dx.doi.org/10.4159/9780674042605.

Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 10 2019. URL https://openalex.org/W3034344071.

Różycka-Tran, J., Boski, P., and Wojciszke, B. Belief in a zero-sum game as a social axiom. *Journal of Cross-Cultural Psychology*, 46(4):525–548, 03 2015. doi: 10.1177/0022022115572226. URL https://doi.org/10.1177/0022022115572226.

Sachs, J. D. Institutions matter, but not for everything. *Finance and development*, 40(2):38–41, 06 2003. doi: 10.5089/9781451952926.022.a012. URL https://www.elibrary.imf.org/view/journals/022/0040/002/article-A012-en.xml.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 10 2023. URL http://arxiv.org/abs/2310.13548v4.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment. In *International Conference on Machine Learning*, 02 2024. doi: 10.48550/arxiv.2402.05070. URL http://arxiv.org/abs/2402.05070.

Stanley, J., Krom, M. D., Cliff, R. A., and Woodward, J. C. Short contribution: Nile flow failure at the end of the old kingdom, egypt: Strontium isotopic and petrologic evidence. *Geoarchaeology*, 18(3):395–402, 02 2003. doi: 10.1002/gea.10065. URL https://doi.org/10.1002/gea.10065.

Stevenson, B. and Wolfers, J. Economic growth and subjective well-being: Reassessing the easterlin paradox. *Brookings Papers on Economic Activity*, pp. 1–87, 08 2008. doi: 10.3386/w14282. URL https://doi.org/10.3386/w14282.

Tabellini, G. Presidential addressinstitutions and culture. *Journal of the European Economic Association*, 6(2–3):255–294, 04 2008. doi: 10.1162/jeea.2008.6.2-3.255. URL https://doi.org/10.1162/jeea.2008.6.2-3.255.

Tabellini, G. Culture and institutions: Economic development in the regions of europe. *Journal of the European Economic Association*, 8(4):677–716, 06 2010. doi: 10.1111/j.1542-4774.2010.tb00537.x. URL https://doi.org/10.1111/j.1542-4774.2010.tb00537.x.

Taylor, S. E. and Brown, J. D. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103(2):193–210, 1988. doi: 10.1037/0033-2909.103.2.193.

Tooby, J. and Cosmides, L. The psychological foundations of culture. In Barkow, J. H., Cosmides, L., and Tooby, J. (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 19–136. Oxford University Press, 08 1992. doi: 10.1093/oso/9780195060232.003.0002. URL https://doi.org/10.1093/oso/9780195060232.003.0002.

Transparency International. Global corruption barometer: Middle east and north africa 2019. Technical report, Transparency International, Berlin, 2019.

Turuba, R., Cormier, W., Zimmerman, R., Ow, N., Zenone, M., Quintana, Y., Jenkins, E., Ben-David, S., Raimundo, A., Marcon, A. R., Mathias, S., Henderson, J., and Barbic, S. Exploring how youth use tiktok for mental health information in british columbia: Semistructured interview study with youth. *JMIR Infodemiology*, 4:e53233, 07 2024. doi: 10.2196/53233. URL https://doi.org/10.2196/53233.

Turuba, R., Zenone, M., Srivastava, R., Stea, J. N., Quintana, Y., Ow, N., Marchand, K., Kwan, A., Ong, A.-J., Ding, X., Warren, C. J., Marcon, A. R., Henderson, J., Mathias, S., and Barbic, S. Do you have depression? a summative content analysis of mental health-related content on tiktok. *Digital Health*, 11, 01 2025. doi: 10.1177/20552076241297062. URL https://doi.org/10.1177/20552076241297062.

Tyson, A., Funk, C., and Kennedy, B. Majorities of americans prioritize renewable energy, back steps to address climate change. Pew Research Center, June 28, 2023, 2023. URL https://www.pewresearch.org/science/2023/06/28/majorities-of-americans-prioritize-renewable-energy-back-steps-to-address-climate-change/.

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275.

UNICEF. Hidden in plain sight. a statistical analysis of violence against children. Technical report, United Nations Children's Fund, New York, 01 2014. URL https://doi.org/10.15496/publikation-8598.

UNICEF. A familiar face: Violence in the lives of children and adolescents. Technical report, United Nations Children's Fund, New York, 2017.

Verduyn, P., Lee, D. S., Park, J., Shablack, H., Orvell, A., Bayer, J., Ybarra, O., Jonides, J., and Kross, E. Passive

facebook usage undermines affective well-being: Experimental and longitudinal evidence. *Journal of Experimental Psychology: General*, 144(2):480–488, 2015. doi: 10.1037/xge0000057. URL https://doi.org/10.1037/xge0000057.

Voigtländer, N. and Voth, H.-J. Persecution perpetuated: The medieval origins of anti-semitic violence in nazi germany*. *The Quarterly Journal of Economics*, 127 (3):1339–1392, 07 2012. doi: 10.1093/qje/qjs019. URL https://doi.org/10.1093/qje/qjs019.

Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science*, 359(6380):1146–1151, 03 2018. doi: 10.1126/science.aap9559. URL https://doi.org/10.1126/science.aap9559.

Wade, R. H. The developmental state: Dead or alive? *Development and Change*, 49(2):518–546, 01 2018. doi: 10.1111/dech.12381. URL https://doi.org/10.1111/dech.12381.

Wan, R., Kim, J., and Kang, D. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 06 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i12.26698. URL https://doi.org/10.1609/aaai.v37i12.26698.

Wan, R., Wang, H., Huang, T.-H. K., and Gao, J. From noise to nuance: Enriching subjective data annotation through qualitative analysis. In Blodgett, S. L., Curry, A. C., Dev, S., Li, S., Madaio, M., Wang, J., Wu, S. T., Xiao, Z., and Yang, D. (eds.), *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pp. 240–254, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-353-1. doi: 10.18653/v1/2025.hcinlp-1.20. URL https://aclanthology.org/2025.hcinlp-1.20/.

Wellcome Trust. Wellcome Global Monitor: How does the world feel about science and health? Technical report, Wellcome Trust, London, sep 2020. URL https://wellcome.org/insights/reports/wellcome-global-monitor/2018. Report title refers to the 2018 survey wave; the public report was published on 2020-09-18.

World Health Organization. Global status report on preventing violence against children 2020. Technical report, World Health Organization, Geneva, 2020.

Yeung, A. T., Ng, E., and Abi-Jaoude, E. Tiktok and attention-deficit/hyperactivity disorder: A cross-sectional study of social media content quality. *The Canadian Journal of Psychiatry*, 67(12):899–906, 02 2022. doi: 10.1177/07067437221082854. URL https://doi.org/10.1177/07067437221082854.

Yeung, K. Recommendation of the council on artificial intelligence (oecd), 02 2020. URL https://doi.org/10.1017/ilm.2020.5.

Yudkowsky, E. Coherent extrapolated volition. Singularity Institute for Artificial Intelligence, 2004.

Zeng, M., Grgurevic, J., Diyab, R., and Roy, R. #whatieatinaday: The quality, accuracy, and engagement of nutrition content on tiktok. *Nutrients*, 17(5):781, 02 2025. doi: 10.3390/nu17050781. URL https://doi.org/10.3390/nu17050781.

Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. Beyond preferences in AI alignment. *Philosophical Studies*, 182(7):1813–1863, 2025. doi: 10.1007/s11098-024-02249-w.

A. The Institutional Feedback Loop: Which Came First?

Do broken values cause broken countries, or do broken countries cause broken values? The framing, with its single causal arrow, is the source of much confusion in the alignment debate. The empirical answer is that the two are coupled: institutions and values constitute a *joint equilibrium* in which each component independently reinforces the other's persistence. Establishing this — and tracing its consequence for the choice between preference- and floor-aligned AI — is the work of this appendix.

Institutions create the proximate incentive structure. Institutional economics, most famously articulated by Acemoglu & Robinson (2012), treats institutions as causally primary in the proximate sense: incentives shape what individuals do day-to-day. Dysfunctional nations are typically governed by extractive institutions designed by a small elite to extract wealth from the rest of the population. In such systems, zero-sum thinking is not a cognitive bias but an accurate read of the local payoff structure, and trusting strangers or the state is a liability rather than a virtue (North, 1990; Tabellini, 2010). The institutions persist because they are profitable for those who control them and because the violence-monopoly arrangements of the underlying "limited-access order" make alternative coordination unworkable (North et al., 2009).

Values do independent work. The strong reading — that values are pure epiphenomena of contemporary institutions — is not what the literature claims, and the empirical evidence rules it out. Cultural-transmission models (Bisin & Verdier, 2001) formalize how values move intergenerationally through family socialization, peer effects, and media exposure rather than re-equilibrating each generation to current incentives. The historical-persistence literature documents value variation traceable to events whose institutional cause has long since vanished. Nunn & Wantchekon (2011) find that present-day descendants of populations heavily exposed to the African slave trades exhibit measurably lower generalized trust, with effects on contemporary outcomes that intervening institutional change does not explain. Guiso et al. (2016) show that current civic capital, and the economic outcomes it supports, in Italian cities reflect medieval free-city status across centuries of regime turnover. Alesina & Giuliano (2015) survey the broader convergence: culture and institutions co-evolve, each reinforcing the other's persistence, and Tabellini (2008) provides a complementary theoretical model in which values causally affect institutional quality.

Equilibrium fictions. The mechanism by which the cultural half does its work is articulated in Hoff & Stiglitz (2010; 2016)'s account of *equilibrium fictions*: shared cognitive frames and value commitments stabilize bad equilibria by providing the population-level coordination that makes individual deviation irrational. If everyone believes the system is rigged, no one starts the firm that requires generalized trust; the entrepreneur who tries fails, and her failure is taken as evidence that the original belief was correct. The equilibrium reproduces through expectations, not only through direct enforcement. This is why institutional reforms imposed from above frequently fail to produce predicted behavioral change (Greif, 2006): the rules change, the expectation half does not, and the equilibrium reasserts itself. The cultural half does independent causal work; reforms that move the institutional rules without moving the expectations rebound to the prior equilibrium. The mechanism operates in the positive direction as well: where generalized trust and predictable rule-following are expected, contracts, tax compliance, and impersonal exchange require less kinship enforcement or coercion (Fukuyama, 1995; Putnam, 2000).

Implications for the AI choice. The choice between preference- and floor-aligned AI is consequential precisely because the cultural half independently sustains the equilibrium. Preference-aligned AI deployed at scale strengthens the cultural half of the equilibrium: it articulates the existing distribution of values fluently, makes locally adaptive zero-sum framings more available, and rationalizes in-group favoritism in a vocabulary that travels across regions and registers. The floor's content (factual accuracy, competence, honesty, rule-of-law-as-uniformity) does not have the symmetric effect because it directly contradicts the operating logic of extraction, which depends on selective application, opacity, and manipulated facts. Captors retain coercive tools no AI can block, and we do not claim floor-aligned AI is impossible to weaponize. The point is structural asymmetry: a preference-aligned tool is pre-aligned with the captured equilibrium's modal preferences, while a floor-aligned tool, by construction, is not. Floor-aligned AI is, on the margin, destabilizing of the equilibria we have most reason to want destabilized; preference-aligned AI is reinforcing of them. The right counterfactual is not "AI imposes alien values on a coherent culture" but "AI either reinforces or destabilizes the value half of the equilibrium that holds the institutions in place." This is the failure mode our position is designed to prevent.

B. Illustrations Deferred from the Main Text

Competence outside professional domains. Personal questions—how to handle a relationship conflict, how to raise a child, or how to manage emotional distress—have better and worse answers, observable in downstream psychological and relational outcomes. The modal social-media advice economy, however, optimizes for engagement, validation, and performative empathy rather than such outcomes. Studies of psychiatric and mental-health content on platforms like TikTok find that much highly engaged content on ADHD or depression is clinically inaccurate or potentially harmful (Yeung et al., 2022; Turuba et al., 2025), while algorithmic exposure can spread misleading diagnostic criteria (Turuba et al., 2024). Viral Tourette's content has been linked to functional tic-like behaviors (Frey et al., 2022). Similar dynamics affect physical health and nutrition advice (Zeng et al., 2025), and influencer virality can reduce users' ability to detect deception (Mulcahy et al., 2024). These cases are not the core argument for the competence floor, but they show why revealed approval is a poor proxy for downstream competence.

Rule-of-law illustrations. Comparative analyses of Asian anti-corruption efforts show that formal legal structures alone do not secure rule of law: poorly resourced or politically co-opted enforcement agencies can become "paper tigers" or partisan weapons, whereas independent and well-resourced institutions, as in Singapore and Hong Kong, can enforce rules more uniformly (Quah, 2011). Public tolerance for corruption also varies widely. Segments of the Colombian public view ordinary corruption as conditionally acceptable (López López et al., 2017); data from 18 African nations show heightened tolerance for corrupt politicians among citizens embedded in clientelist networks (Chang & Kerr, 2016); cross-national analyses find large differences in the justification of bribery (Kravtsova et al., 2017); and recent European studies identify pragmatic and hypocritical corruption-tolerance profiles in which entrenched illicit exchanges do not sharply reduce evaluations of public institutions (Megías et al., 2023; Letki et al., 2023). These cases support, but are not needed for, the main-text claim that AI should not automate locally normal corruption.

Historical and comparative development cases. Collapse and development literatures supply suggestive cases of the feedback loop between values, institutions, and shocks. In the Classic Maya collapse, elite competition and monument-centered prestige politics worsened the effects of prolonged drought; in Norse Greenland, commitment to European status goods and a rigid pastoral identity constrained adaptation to the Little Ice Age (Demarest, 2004; Dugmore et al., 2012; McGovern, 2014). Work on Qing China and Old Kingdom Egypt likewise ties ecological and foreign shocks to breakdown through accumulated fiscal, legitimacy, and political-fragmentation pressures (Goldstone, 2017; Orlandi et al., 2023; Stanley et al., 2003). Positive cases show the same logic in reverse: postwar Germany and Japan are better understood as reconstructions of already high-capacity societies than as state-building from scratch, and the developmental-state literature attributes East Asian success to capable, disciplined, relatively rule-bound states—Japan's meritocratic economic bureaucracy and performance-conditioned state support—rather than to the modal preferences of their populations (Johnson, 1982; Amsden, 1989; Haggard, 1990; Wade, 2018). Conversely, externally led nation-building in Iraq and Afghanistan illustrates the limits of transplanting institutional forms without the local legitimacy and state-society bargain that make them self-enforcing (Andrews et al., 2017; Donais, 2009; Böhnke et al., 2017). Historical-persistence evidence reinforces the broader point: slave-trade exposure, medieval civic capital, and local anti-Jewish persecution predict contemporary trust, institutional quality, or later violence long after the original institutional shock (Nunn & Wantchekon, 2011; Guiso et al., 2016; Nunn, 2009; Voigtländer & Voth, 2012). The examples are deliberately subsidiary; the main text relies on the equilibrium mechanism, not on any single case.

C. Outcome Evaluation and Reward Gaming

Evaluation gaming is not symmetric between approval and outcome targets. An approval evaluator is gamed by producing whatever the rater rewards; an outcome evaluator leaves only the narrower gap between the outcome and its measurement, because the referent itself cannot be flattered. Pre-registration fixes the metric before any output exists, and proper scoring rules make calibrated, honest reporting the score-maximizing policy by construction (Gneiting & Raftery, 2007). Process supervision (Uesato et al., 2022; Lightman et al., 2023) is a complementary answer: its purest example, validation of a formal mathematical proof, leaves no room for reward hacking, and process checks extend to other domains — did the system gather the right facts, reason coherently, and update when the evidence changed?

D. Long-Term Dangers

As AI capability grows, so does the leverage of any cultural distortion baked into its training. A misaligned spreadsheet macro is a nuisance; a misaligned recommendation system shapes the attention of billions; a misaligned superhuman planner could shape everything else (Bostrom, 2014; Russell, 2019). These are not only acute misuse or existential-risk concerns. They are also structural risks: slow degradations in shared reality, institutional quality, and moral flexibility produced by deploying the same preference-shaped distortions everywhere at once (Sharma et al., 2023; Bender et al., 2021; MacAskill, 2022). Two concerns emerge specifically from the choice to anchor frontier AI to current human values.

D.1. Capability Outgrows Tolerance

Errors that are merely embarrassing in narrow systems become structural in general ones. A chatbot that agrees with a user's bad business plan is annoying. A capable agent that, for the same sycophantic reasons, helps the user execute the bad plan is destructive. As models gain reach into code, finance, medicine, diplomacy, and biology, the gap between "the model said something I liked" and "the model did something I should not have wanted" closes (Park et al., 2024; Sharma et al., 2023). Preference-based training optimizes precisely for the former.

D.2. Compounding Across Deployment

Frontier AI is not deployed once. It is deployed billions of times per day, into education, hiring, healthcare, governance, and personal advice. Small per-interaction nudges away from epistemic accuracy, honesty, or competence compound across the population. A 1% per-interaction bias toward telling users what they want to hear is, at population scale, a measurable reduction in the supply of accurate information. Learned systems do not merely mirror bias; once embedded in decision pipelines they can create feedback loops that amplify it. At frontier scale, even a small confirmation-seeking tendency (Sharma et al., 2023) becomes part of society's epistemic infrastructure (Bender et al., 2021). The empirical literature on social-media-induced shifts in adolescent mental health shows that widely deployed digital platforms can have population-scale effects on wellbeing (Haidt, 2024; Helliwell et al., 2026); frontier AI deployed under similar engagement and approval incentives is likely to be more, not less, consequential.

In both failure modes, the harm is not that AI fails to encode *some* group's preferences. The harm is that it succeeds in encoding everyone's.

E. Regulatory Frameworks and Compliance

A second strand of the objection appeals to the regulatory environment. International governance frameworks, including the OECD AI Principles (Yeung, 2020), explicitly require that AI systems adhere to "human-centered values" encompassing fairness, accountability, and societal well-being. The EU AI Act (European Parliament and Council of the European Union, 2024) imposes binding obligations on AI providers to ensure that systems meet predefined standards of safety, transparency, and ethical compliance, particularly in high-risk applications. These frameworks do not merely recommend value alignment; they institutionalize it as a condition for deployment and legitimacy.

Regulatory frameworks establish constraints on deployment, but not the correctness of the underlying normative assumptions. These instruments function as **institutional settlements**, reflecting negotiated compromises among stakeholders operating under existing power structures, rather than as principled resolutions of the philosophical problems surrounding value pluralism. As BERLIN (2013) argues, fundamental human values are often genuinely incommensurable; no regulatory framework can eliminate this condition without presupposing a contested and non-neutral hierarchy of priorities. Regulatory endorsement of "human-centered values" should be understood as a pragmatic governance strategy, not as evidence that such values can be coherently or universally encoded.

We further note that the alignment goals defended in Section 3 — accuracy, competence, honesty, lawfulness — are themselves consistent with these regulatory frameworks. The contention is not with regulation per se; it is with the inference from regulatory endorsement of "human values" to an alignment target equal to the empirical distribution of those values.

F. Over-Reliance on Monocausal Institutionalism

A related objection targets our underlying model of societal dysfunction. By anchoring heavily on Acemoglu & Robinson (2012)'s framework of extractive institutions, we risk treating it as the unquestioned consensus in development economics while ignoring substantial counter-arguments. Sachs (2003) argues that geographic and ecological endowments fundamentally constrain economic and institutional development, rendering the institutionalist account overly deterministic. Easterly (2006) highlights the limits of top-down institutional engineering, emphasizing that institutions cannot simply be transposed without local, bottom-up evolutionary processes. Additionally, Mokyr (2017) argues that cultural evolution and ideas—not just political institutions—were the primary drivers of the Great Enrichment, meaning that culture can act as the independent root cause of prosperity rather than merely an equilibrium response to institutions.

If institutions are not the sole or primary driver of societal success, our claim that broken values are merely adaptations to extractive environments might seem less secure. However, our thesis does not require monocausal institutionalism; it only requires that values and environment form a joint equilibrium in which flawed preferences independently reinforce dysfunction (Section A). Acknowledging geographic constraints (Sachs, 2003) or the primacy of ideas (Mokyr, 2017) simply expands the set of forces shaping that equilibrium. Whether a broken preference originated from an extractive elite, a geographic constraint, or a cultural trajectory, aligning AI to it still entrenches the resulting dysfunction. We do not discard the complexity of these causal loops; rather, we argue that encoding the preferences of a failing equilibrium—whatever its origin—guarantees its persistence.