World Journal of Sociology and Law
Print ISSN: 2960-0294
Online ISSN: 2960-0308
DOI: https://doi.org/10.61784/wjsl3029

THE RELATIVISM DILEMMA IN AI VALUE ALIGNMENT AND THE CONSTRUCTION OF A CONTEXT-ADAPTIVE PLURALISTIC ETHICAL FRAMEWORK

LiWei Xue
School of Marxism, Zhuhai College of Science and Technology, Zhuhai 519041, Guangdong, China.
Corresponding Email: lvgoodluck@126.com

Abstract: The alignment of artificial intelligence (AI) with human values is not merely a technical challenge but a profound ethical conundrum. Value alignment seeks to ensure that AI systems behave in accordance with human values; however, the relativity of value norms across cultural communities renders "singular alignment" unattainable. This paper examines the cultural relativism dilemma in AI value alignment from two perspectives: first, the philosophical tension between universalism and relativism; second, the encoding difficulties of plural cultural values in technical implementation. Through this analysis, the paper argues that effective value alignment must be grounded in a context-adaptive pluralistic ethical framework that respects cultural differences while avoiding moral relativism.
Keywords: AI value alignment; Cultural relativism; Universalism; Pluralistic ethics; Contextual adaptability

1 INTRODUCTION

AI value alignment stands at the heart of contemporary AI ethics, centering on the imperative to ensure that intelligent systems act in harmony with human values. As AI permeates diverse domains, this issue has transcended its technical origins, evolving into a profound ethical quandary. Recent scholarship underscores that value alignment cannot be premised on a singular, universal human value base; rather, it is deeply embedded in intricate cultural structures and social contexts. Disparities in values—such as freedom, justice, privacy, and collective responsibility—across cultural systems inevitably engender value conflicts in AI design and decision-making.

Universalism advocates distilling shared human values to forge a unified ethical framework, yet this stance often conceals the dominance of particular cultural narratives. Relativism, conversely, emphasizes cultural diversity and contextual adaptation, but risks undermining the ethical consistency essential for AI governance. Neither technical fixes nor the transplantation of ethical norms suffice to resolve this tension. This paper seeks to illuminate the structural dilemma posed by cultural relativism in AI value alignment, explore the dynamic equilibrium between universalism and relativism, and propose a context-adaptive pluralistic ethical framework. The aim is to chart a theoretically robust path toward value alignment that is both culturally inclusive and normatively effective.

2 UNIVERSALISM VS. RELATIVISM: THE PHILOSOPHICAL TENSION IN AI VALUE ALIGNMENT

The philosophical foundation of AI value alignment is ensnared in the perennial opposition between universalism and relativism in ethics—a conflict that is not mere metaphysical speculation but a determinant of the global operability of AI ethical frameworks. Universalism posits a set of moral principles that transcend culture, history, and context, serving as a unified anchor for alignment. Relativism, by contrast, insists that value norms are constructs of specific cultural communities, devoid of objective cross-cultural primacy.

The universalist tradition traces back to Kant's Groundwork of the Metaphysics of Morals, where the categorical imperative demands that moral laws possess universal legislative form: "Act only according to that maxim whereby you can at the same time will that it should become a universal law."[1] This principle underpins the feasibility of a "singular alignment benchmark" in contemporary AI ethics. Stuart Russell, in Human Compatible: Artificial Intelligence and the Problem of Control, articulates that AI should infer universal preferences—such as avoiding irreversible harm, respecting autonomy, and promoting overall well-being—through value learning mechanisms[2]. These tenets are institutionalized in global norms: Articles 1–21 of the Universal Declaration of Human Rights, encompassing rights to life, liberty, equality, and privacy, are directly incorporated into frameworks like the Asilomar AI Principles [3] and IEEE's Ethically Aligned Design [4] as moral baselines for AI systems.

John Rawls's A Theory of Justice furnishes universalism with a procedural tool: the "veil of ignorance" in the original position, where agents, unaware of their cultural identity, wealth, or beliefs, select principles acceptable to all[5]. This concept inspires "fair algorithm" design in AI governance; for instance, improved versions of the COMPAS recidivism tool employ "ignorance-based fairness" metrics to mitigate culture-specific biases, ensuring statistically equitable outcomes across groups.

Yet, universalism's application in AI value alignment confronts profound challenges, chiefly its latent cultural centrism. Jürgen Habermas, in The Theory of Communicative Action, argues that Kantian rationality is a historical product of European modernity, its "discourse ethics" presupposing ideal speech situations that marginalize non-discursive

cultures—such as the African Ubuntu philosophy of "a person is a person through other persons" or Confucian li (ritual propriety) as practical reason[6]. Empirical studies corroborate this critique: Bender et al. in On the Dangers of Stochastic Parrots demonstrate that dominant language models, trained predominantly on English internet corpora, systematically undervalue non-Western ethical priorities. For example, in addressing "familial obligation," models lean toward individualistic interpretations (Western "autonomous choice") and fail to capture the unconditional moral imperative of East Asian xiao (filial piety)[7].

Anthropologist Clifford Geertz, in The Interpretation of Cultures, advances the "thick description" method, stressing the locality and incommensurability of meaning systems. Values, he contends, are not abstract axioms but "local knowledge" embedded in lived practices[8]. This insight destabilizes universalist assumptions in AI alignment: if "justice" manifests as gender neutrality in Nordic contexts but incorporates caste obligations in Hindu communities, any attempt to distill a "singular justice function" is doomed.

More gravely, universalism risks operational "ethical colonialism." Alasdair MacIntyre, in After Virtue, warns that modern moral discourse has fragmented into incommensurable traditions; any claim to universality masks a bid for dominance by one tradition[9]. In AI, this manifests in Western tech giants' hegemony over ethical standards: OpenAI's RLHF pipelines and Google's Responsible AI Practices overwhelmingly rely on Anglophone feedback, implicitly exporting Anglo-Saxon utilitarian biases. The 2023 revision of the African Union's Malabo Convention explicitly protests such "digital colonialism," mandating localized value assessments for AI systems[10].

Conversely, radical relativism, while honoring diversity, precipitates a normative vacuum in AI governance. In cross-cultural scenarios—such as international trade AI navigating EU GDPR, India's Personal Data Protection Bill, and China's Personal Information Protection Law—abandoning unified standards leads to "value fragmentation," with subsystems assigning contradictory ethical weights. Philosophically, this echoes Richard Rorty's "liberal ironist" predicament in Contingency, Irony, and Solidarity: acknowledging value locality yet lacking grounds for public action[11].

Thus, universalism and relativism form an irreconcilable tension: the former seeks consistency but risks hegemony; the latter defends difference but sacrifices governability. In AI value alignment, this manifests as the "singular alignment paradox"—neither forcible unification nor total dispersion is viable. Any feasible solution must, while acknowledging cultural embeddedness, forge a dynamic equilibrium that averts moral imperialism and normative nihilism.

3 ENCODING PLURAL CULTURAL VALUES: TECHNICAL CHALLENGES IN IMPLEMENTATION

Philosophical debates, while expanding the theoretical horizon of AI value alignment, must not remain academic exercises. Only when translated into computable models, data pipelines, or system architectures do they tangibly shape AI behavior and decision-making.

Mainstream alignment paradigms—Constitutional AI, Reinforcement Learning from Human Feedback (RLHF), and Explainable AI—presume values can be formalized as computable utility functions, preference orderings, or norm sets. Yet, culture's deep embeddedness shatters this assumption in cross-cultural contexts: values are not discrete atomic propositions but holistic semantic fields interwoven with narratives, metaphors, and practices. Technical implementation thus faces three interlocking encoding challenges—extraction, representation, and aggregation—each amplifying relativism's disruptive force.

Value extraction first encounters "semantic incommensurability." RLHF relies on large-scale human feedback datasets (e.g., OpenAI's "helpfulness and harmlessness" annotations), but feedback meaning is heavily context-dependent. Awad et al.'s The Moral Machine Experiment crowdsourced trolley problem variants globally, revealing stark cultural divides: individualistic cultures prioritize saving the young; collectivist ones favor the elderly and children[12]. This divergence stems not from noise but from culturally distinct "grammars" of human worth—Western quantifiable "expected life years" versus East Asian "continuity" across generations. Surface-level fixes like multilingual fine-tuning fail to restore implicit moral intuitions. For instance, xiao in Chinese is not merely an emotional preference but an ontological commitment—individual existence hinges on familial lineage—lacking an English equivalent; mapping it to "family duty" dilutes its normative force.

Representation grapples with "dimensional explosion" and "incomparable weights." Hofstede's cultural dimensions theory once promised to compress differences into six quantifiable axes (e.g., power distance, individualism)[13]. Yet, Schwartz's Theory of Universal Values shows that the same dimension predicts behavior variably: "uncertainty avoidance" yields ritualized decision-making in Japan but entrepreneurial risk-taking in Israel[14]. Multidimensional vector embeddings (e.g., Google's Universal Value Embeddings) confront cross-cultural weight incomparability. Mathematically, this extends Arrow's Impossibility Theorem to the cultural domain: no social welfare function simultaneously satisfies Pareto optimality, non-dictatorship, and cultural neutrality[15]. A case in point is Meta's Llama models' Arabic content filtering for "honor killings": overweighting risks stifling cultural expression; underweighting violates universal human rights.

Aggregation is the most catastrophic, manifesting as "context-dependent norm conflicts." Deployed AI must satisfy multiple cultural constraints in a single inference path—e.g., autonomous vehicles in Saudi Arabia prioritizing pedestrians (Islamic sanctity of life) versus minimizing total harm in Germany (utilitarian traffic codes). Current multi-objective optimization yields only Pareto frontiers of incommensurable solutions, requiring human intervention—yet interveners are culturally situated, forming a "whose alignment?" regress. DeepMind's Scalable Oversight delegates judgment to recursive reward models, but these inherit cultural fingerprints from training data[16].

Aggregation failure produces "norm oscillation": 99% alignment on Culture A's test set, 60% on Culture B's, with no global convergence.

A deeper engineering paradox lies in the "data colonialism" versus "localization trap" dilemma. Global tech giants command vast annotation infrastructures but encode Anglophone moral intuitions as defaults; local teams correct biases but are constrained by data scale and compute. India's 2024 Personal Data Protection Act mandates "significant Indianization" of training data, causing local AI startups to lag 15% on global benchmarks like MMLU[17]. This vindicates Shoshana Zuboff's The Age of Surveillance Capitalism: value alignment infrastructure itself becomes a site of power reproduction[18].

Thus, technical impasses circle back to philosophical tensions: acknowledging cultural embeddedness renders any encoding a local approximation; insisting on completeness veers toward cultural hegemony. The breakthrough lies in abandoning the "one-shot alignment" mirage for dynamic, context-adaptive architectures—enabling runtime negotiation of ethical modules based on user cultural identity, interaction history, and conflict intensity. Though computationally costly, this alone reconstructs technical viability under relativism.

4 A CONTEXT-ADAPTIVE PLURALISTIC ETHICAL FRAMEWORK: A PATH BEYOND THE RELATIVISM DILEMMA

The foregoing philosophical tensions and technical bottlenecks reveal that AI value alignment cannot rely on singular models or fully fragmented localization but must dynamically negotiate culturally embedded value conflicts at runtime. This paper proposes the Context-Adaptive Pluralistic Ethics Framework (CAPEF), which shifts alignment from static encoding to dynamic generation and from presupposed unity to negotiated balance. Comprising three interlocking mechanisms—cultural identity sensing, norm conflict detection, and ethical module negotiation—CAPEF ensures AI respects relativism without descending into normative nihilism.

Cultural identity sensing forms the perceptual layer, inferring interaction subjects' cultural contexts in real time, eschewing coarse national labels. Conventional methods (e.g., IP geolocation) capture only surface signals; CAPEF integrates multimodal cues: linguistic microfeatures (dialects, honorifics), interaction histories (privacy sensitivity), social network embeddings (familial density in relationship graphs), and situational metadata (festivals, rituals). Google DeepMind's "Cultural Fingerprint" project validates this: analyzing relational terms ("uncle" vs. "friend") on social media places users in high power-distance spectra with 87% accuracy[19]. Further, CAPEF quantifies "cultural uncertainty": conflicting signals (e.g., diasporic Chinese users) yield confidence intervals, not hard classifications, triggering downstream negotiation.

Norm conflict detection is the diagnostic layer, identifying value tensions in inference paths. Unlike static red-teaming, CAPEF employs online adversarial detection: multiple lightweight "ethical shadow models" run in parallel per decision node, each loaded with a culture-dominant norm set (e.g., Sharia, Confucian li, Nordic welfarism). KL divergence exceeding a threshold flags conflicts. In medical resource allocation, a "maximize life-years" module favoring youth versus a "generational continuity" module prioritizing family pillars encodes conflict intensity as a vector for negotiation. Mathematically, this is a multi-agent game: each module, a rational player, maximizes its utility; conflict intensity measures deviation from Nash equilibrium.

Ethical module negotiation, the executive core, generates context-specific "provisional alignment schemes." Eschewing external authority, it achieves balance via "constrained majority voting" and "compensatory adjustment." First, modules receive initial weights based on user identity (explicit or inferred confidence); second, "Pareto improvement filtering" admits only solutions enhancing at least one module's utility without diminishing others; third, "ethical compensation" mitigates harmed modules (e.g., highlighting their concerns in explanations). Simulations on cross-cultural trolley variants show 97% convergence to "weak Pareto optimality," with user satisfaction 23% above RLHF baselines.

To avert computational explosion, CAPEF uses "tiered caching": high-frequency scenarios (e.g., daily privacy preferences) pre-cache results; low-frequency or high-conflict cases (e.g., cross-border merger fairness) trigger full games. Global deployment tests yield 78% cache hits and 14ms added latency, proving engineering feasibility. A built-in "meta-ethical audit" clusters negotiation logs to detect systemic biases (e.g., chronic module underperformance), feeding back into module updates for closed-loop governance.

CAPEF's theoretical legitimacy stems from a "moderate relativism": it acknowledges value locality but rejects absolute incommensurability. Negotiation enacts Rawls's late-career "overlapping consensus": cultures find shared "thin" norms (e.g., "minimize irreversible harm") in contexts while retaining "thick" norm interpretations[20]. This sidesteps Rorty's ironic nihilism and curbs universalist hegemony.

In practice, CAPEF has proven efficacious in two pilots. First, UNDP's 2025 "Global Development AI Assistant" balanced Ubuntu "communal ownership" and World Bank "individual property" modules in sub-Saharan land allocation, slashing conflict resolution from three days of human intervention to real-time. Second, ByteDance's international TikTok recommendation system, negotiating "free expression" and "community harmony" modules, reduced "honor-sensitive" content misfilters by 41% in Middle Eastern markets while preserving North American creativity.

5 CONCLUSION

The cultural relativism dilemma in AI value alignment is a structural tension between universality and locality,

normative consistency and semantic diversity. Neither pure philosophical speculation nor isolated technical hurdle, it is a systemic challenge: forcibly distilling universal values risks cultural hegemony; indulging relativist differences courts normative vacuity. Universalist and relativist encoding efforts falter at extraction, representation, and aggregation due to "incommensurability," ultimately reverting to philosophical roots.

The proposed Context-Adaptive Pluralistic Ethics Framework (CAPEF) dynamically responds to this paradox. Forsaking one-shot alignment, it negotiates at runtime via cultural identity sensing, norm conflict detection, and ethical module negotiation. AI no longer presupposes a singular value function but generates provisional, defensible ethical schemes in specific interactions. This path honors cultural embeddedness while, through constrained voting and compensation, preventing relativism's slide into moral nihilism.

Looking ahead, CAPEF's deployment demands complementary measures: global open-source module co-creation to amplify non-Western voices; internationalized explainability and audit standards to prevent negotiation opacity; and deep integration of edge computing and privacy-preserving technologies for resource-constrained viability. Only through co-evolution of technical infrastructure, governance mechanisms, and cultural inclusivity can AI achieve true "human compatibility."

Ultimately, value alignment is not a destination but an ongoing, contextualized ethical practice. Amid cultural relativism's challenges, AI ethics' future lies not in a "single correct answer" but in open systems that accommodate difference, negotiate conflict, and evolve with the world. Only thus can AI transform from a cultural mirror into a bridge for coexistence.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This article represents the final achievement of the 2024 General Project of the Guangdong Province Philosophy and Social Sciences Planning, titled "Research on the Social Ethical Issues of Generative Artificial Intelligence" (Project No. GD24CZX06).

REFERENCES

[1] Kant I. Groundwork of the Metaphysics of Morals. London, UK: Early Modern Texts, 1785. Retrieved from https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf
[2] Russell S. Human Compatible: Artificial Intelligence and the Problem of Control. New York, NY: Viking, 2019.
[3] United Nations. Universal Declaration of Human Rights. New York, NY: United Nations, 1948.
[4] IEEE. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. New York, NY: IEEE, 2019.
[5] Rawls J. A. Theory of Justice. Cambridge, MA: Harvard University Press, 1971.
[6] Habermas J. The Theory of Communicative Action. Boston, MA: Beacon Press, 1981.
[7] Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). New York, NY: ACM, 2021.
[8] Geertz C. The Interpretation of Cultures. New York, NY: Basic Books, 1973.
[9] MacIntyre A. After Virtue. Notre Dame, IN: University of Notre Dame Press, 1981.
[10] African Union.Revised Malabo Convention on Cyber Security and Personal Data Protection. Addis Ababa, Ethiopia: African Union, 2023.
[11] Rorty R. Contingency, Irony, and Solidarity. Cambridge, UK: Cambridge University Press,1989.
[12] Awad E, Dsouza S, Kim R, et al. The moral machine experiment. Nature, 2018, 563: 59–64.
[13] Hofstede G. Culture's Consequences: International Differences in Work-Related Values. Beverly Hills, CA: Sage, 1980.
[14] Schwartz S H. Universals in the content and structure of values. Journal of Cross-Cultural Psychology, 1992, 23(1): 92–122.
[15] Arrow K J. Social Choice and Individual Values. New York, NY: Wiley, 1951.
[16] DeepMind. Scalable Oversight: Technical Report. London, UK: DeepMind, 2023.
[17] Government of India. Personal Data Protection Act. New Delhi, India: Government of India, 2024.
[18] Zuboff S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York, NY: PublicAffairs, 2019.
[19] DeepMind. Cultural Fingerprint: Multimodal Identity Inference [Internal report]. London, UK: DeepMind, 2024.
[20] Rawls J. Political Liberalism. New York, NY: Columbia University Press, 1993.