Beyond Preference Aggregation: Algorithmic Coloniality and the Case for Relational Pluriversal Alignment

Jingtian Zhang[1] (✉) and Yu Lu[2]

[1] The Experimental High School Attached to Beijing Normal University, Beijing 100032, CHINA arthurzhang50@gmail.com
[2] Zhejiang University, Hangzhou Zhejiang 310058, CHINA aniki.yulu@gmail.com

Abstract. AI alignment is commonly framed as a technical problem of aggregating human preferences into a coherent objective function, assuming that values are commensurable and symmetrically formed. In this paper, we challenge this framing by situating alignment within the structural conditions of coloniality embedded in contemporary socio-technical systems. Using algorithmic coloniality as a diagnostic lens, we analyze how AI infrastructures reproduce global asymmetries through three overlapping dynamics: algorithmic oppression, algorithmic exploitation, and algorithmic dispossession. We argue that these mechanisms destabilize the neutrality of preference aggregation and the assumption of a unified value space. Moving beyond both universalist optimization and essentialist cultural categories, we advance a constructive framework: Relational Pluriversal Alignment. This perspective shifts the focus from identity-based preference profiles to a power-sensitive reorientation of epistemic authority. It operates through three normative commitments: prioritizing systemic exposure over abstract representation, tracking structural vulnerability within infrastructures, and fostering translocal solidarity grounded in shared risk. Ultimately, we reposition AI alignment not as a technical optimization problem, but as a contested site of normative order, epistemic authority, and global justice.

Keywords: Algorithmic Coloniality · AI Alignment · Relational Pluriversal Alignment.

1 Introduction

Artificial Intelligence systems are increasingly framed within the paradigm of alignment, where the central objective is to learn and aggregate human preferences into a unified optimization target [1–4]. This framing assumes that human values are commensurable and that preferences can be treated as neutral inputs to technical systems.

In this paper, we challenge this assumption. We argue that contemporary AI alignment cannot be fully understood as a value-neutral optimization problem, because the formation and representation of preferences are themselves shaped

by historical and structural inequalities embedded in socio-technical systems [5, 6].

We situate this critique within the concept of algorithmic coloniality and related work on data colonialism, which highlight how AI infrastructures reproduce global asymmetries of epistemic authority, labor, and risk distribution [7–9]. From this perspective, AI systems are not neutral aggregators of human values, but active participants in shaping which forms of knowledge and experience become legible within computational systems.

This paper develops a relational interpretation of alignment that shifts attention from preference aggregation to the distribution of epistemic authority and structural exposure within AI systems. Rather than treating alignment as a problem of consensus optimization, we emphasize its role in mediating whose perspectives are operationalized in socio-technical decision-making.

We further argue that this reframing reveals AI alignment to be inseparable from broader questions of global justice and epistemic hierarchy.

2 The Coloniality of Power and Algorithmic Coloniality

2.1 The Coloniality of Power

This section draws on Quijano's concept of the "coloniality of power" [10] as a diagnostic lens for understanding how global hierarchies persist beyond formal colonial rule. Quijano's work suggests that colonial domination did not disappear with political independence but continues to structure modern global society through enduring logics of hierarchy and differentiation. These logics can be observed across multiple interrelated domains, including social classification, labor organization, and knowledge production.

The first domain concerns social classification. Within colonial modernity, race operates not as a biological fact but as a historically produced mechanism of hierarchy that organizes populations into differentiated positions of value and authority, historically positioning Europe as the normative reference point.

The second domain concerns the global organization of labor. Although formal colonial rule has ended, asymmetries in economic structure between center and periphery remain persistent. Building on the dependency tradition [11], many formerly colonized regions continue to occupy positions of primary commodity extraction and low-value production, while capital accumulation remains concentrated in the Global North.

The third domain concerns epistemic authority. Eurocentric knowledge systems have historically been treated as universal, whereas other epistemologies are often rendered local, contextual, or non-scientific [9]. This asymmetry continues to shape what counts as valid knowledge and legitimate forms of reasoning in global knowledge production.

2.2 Coloniality in Socio-Technical Systems

Building on this diagnostic lens, recent scholarship has examined how colonial logics extend into contemporary socio-technical systems, particularly digital infrastructures and AI.

The work on data colonialism [7] highlights how the human experience is increasingly transformed into extractable and commodifiable data. Similarly, research on algorithmic systems has shown that computational infrastructures tend to reproduce historically structured inequalities rather than neutralizing them [8].

From this perspective, AI systems are better understood not as neutral technical artifacts but as infrastructures embedded within global histories of knowledge extraction and labor stratification. Their datasets, model architectures, and deployment contexts are all shaped by uneven global relations of production and epistemic authority. This motivates the analytical perspective developed in the following section, where we examine how these dynamics are manifested in contemporary AI systems.

2.3 Algorithmic Coloniality as a Diagnostic Lens

We use algorithmic coloniality as a diagnostic lens to describe how epistemic, economic, and political hierarchies are reproduced through algorithmic systems. We situate algorithmic systems within broader histories of structural inequality, rather than treating them as isolated technical systems.

Rather than proposing a closed theoretical framework, we use this lens to organize the recurring patterns observed in AI systems across different domains. These patterns can be broadly associated with three interrelated manifestations: algorithmic oppression, algorithmic exploitation, and algorithmic dispossession. These are not discrete categories but analytically overlapping ways in which structural inequalities become operationalized through computational infrastructures.

Algorithmic oppression refers to the reinforcement and amplification of historically sedimented social hierarchies through computational decision-making systems. This builds on analyses of search engines and algorithmic classification as sites of racialized and gendered distortion [6, 12]. It can be observed in systems where classification, ranking, and prediction reproduce structural inequalities under the appearance of neutrality.

Algorithmic exploitation refers to the distributed extraction of labor and value embedded within AI production infrastructures. Drawing on digital labor and platform capitalism literature [13, 14] as well as empirical studies of data work [15], this refers to the global stratification of cognitive and affective labor that sustains AI systems while concentrating value accumulation in dominant technological centers.

Algorithmic dispossession refers to the epistemic and infrastructural reconfiguration of knowledge production under computational regimes. This includes

the privileging of computationally legible forms of knowledge and the marginalization or transformation of alternative epistemologies [7, 9]. It highlights how datafication reshapes what can be known, represented, and recognized within AI systems.

Importantly, these should be understood not as a taxonomy or formal decomposition, but as three overlapping analytical perspectives that together reveal how colonial logics persist within contemporary algorithmic infrastructures. The following section builds on this diagnostic lens to examine how these forms of algorithmic coloniality are instantiated in concrete AI systems and deployment settings.

3 Mechanisms and Instantiations of Algorithmic Coloniality

This section illustrates how algorithmic coloniality manifests in contemporary AI systems through recurring sociotechnical patterns that reproduce structural inequalities in practice.

3.1 Algorithmic Oppression

Algorithmic oppression can be observed when historically produced social hierarchies are encoded and reinforced through automated decision-making systems [12].

In domains such as policing, credit scoring, and content moderation, predictive models are often trained on data that reflect historically unequal social arrangements. Consequently, algorithmic systems can transform structural inequalities into seemingly neutral computational outputs. Predictive policing systems, for instance, use historical arrest records to guide future surveillance allocation, while credit scoring models embed socioeconomic inequalities into assessments of financial risk. In both cases, algorithmic decision-making reproduces existing hierarchies through the appearance of technical objectivity. Similar dynamics have also been observed in administrative governance. In the Dutch childcare allowance scandal (Toeslagenaffaire) [16], algorithmic risk assessment procedures disproportionately targeted immigrant families, including many with dual citizenship linked to former Dutch colonies, illustrating how historically situated categories of difference can become embedded within apparently neutral systems of automated decision-making.

From the perspective of algorithmic coloniality, these systems demonstrate how historical relations of power can be rearticulated through apparently objective computational infrastructures.

3.2 Algorithmic Exploitation

Algorithmic exploitation refers to the extraction of value from human labor that remains essential to AI development while being rendered economically

invisible and systematically undervalued. Although AI is frequently portrayed as a technology of automation, its development relies extensively on human labor. Tasks such as data annotation, content moderation, and Reinforcement Learning from Human Feedback (RLHF) [15] are routinely outsourced to workers in regions such as Kenya, India, and the Philippines, where labor costs are lower and regulatory protections are often weaker.

This pattern is not merely a matter of operational efficiency. Scholars of platform capitalism and digital labor have argued that contemporary digital infrastructures systematically externalize labor costs while concentrating economic value within a small number of technology firms located primarily in the Global North [13, 14]. As a result, the benefits generated by AI systems are often unevenly distributed across global production networks.

RLHF provides a particularly clear example of this dynamic. To improve model safety and alignment, contractors are frequently required to review toxic, violent, or otherwise disturbing content. While these workers bear substantial psychological and emotional risks, the resulting improvements in model performance and commercial value are largely captured elsewhere, revealing an asymmetric distribution of benefits and harms within AI production systems. Empirical reports have documented that such labor is often outsourced to workers in countries such as Kenya, where contractors may be paid less than $2 per hour while performing intensive content evaluation tasks for large-scale AI systems [17]. This illustrates how RLHF pipelines operationalize a global division of labor in which the burdens of model alignment are externalized to economically and geographically marginalized populations.

3.3 Algorithmic Dispossession

Algorithmic dispossession refers to the reconfiguration of epistemic authority through AI systems, whereby certain forms of knowledge become recognized as legitimate while others are rendered invisible, secondary, or unusable. As Mignolo argues, modern epistemic systems are structured by hierarchies that universalize particular forms of knowledge while marginalizing alternative ways of knowing [9].

AI systems can reproduce these hierarchies through processes of datafication and computational standardization. Dataset curation, benchmark design, and alignment procedures determine which forms of knowledge are collected, represented, evaluated, and ultimately incorporated into model behavior. In doing so, they privilege forms of knowledge that are computationally legible while excluding or transforming epistemologies that resist formalization.

From the perspective of algorithmic coloniality, this process extends beyond questions of representation. It concerns the redistribution of epistemic authority itself: whose knowledge becomes encoded into technical systems, whose perspectives shape model behavior, and whose ways of knowing are rendered peripheral within AI-mediated infrastructures.

4 From Universal Alignment to Relational Pluriversal Technoscience

4.1 The Limits of Universal Alignment

Contemporary AI alignment is commonly formulated as the problem of learning and aggregating human preferences into a unified objective function. This framing implicitly assumes that human values are commensurable, that agents are symmetrically situated, and that preference aggregation can be treated as a neutral technical procedure.

However, as argued in the preceding sections, such assumptions obscure the structural conditions under which preferences are formed. Drawing on the concept of algorithmic coloniality, we contend that AI systems are embedded within historically produced asymmetries of knowledge, labor, and authority. These asymmetries shape not only the distribution of data but also the formation of what is treated as "preference" in alignment systems.

From this perspective, universal alignment is not merely a technical aggregation problem but a sociotechnical framing that risks reproducing existing global hierarchies under the appearance of optimization. This calls into question the assumption of a single coherent value space as both theoretically and politically stable.

4.2 The Limits of Existing Pluriversal Alignment Discourse

In response to the limitations of universalist frameworks, recent scholarship in decolonial AI and pluriversal thinking [18, 19] has emphasized the need to recognize multiple epistemologies and value systems. However, we argue that some interpretations of pluriversal alignment [20] remain limited in two respects.

First, they may implicitly rely on relatively fixed representations of cultural or geopolitical categories. By treating groups such as "Indigenous communities", "Global South users" or "Western societies" as stable units of preference [21], such approaches risk reifying the very epistemic boundaries they seek to challenge. This produces a flattened ontology of difference, where internally heterogeneous and contested social formations are reduced to static preference profiles.

Second, such framings can inadvertently reproduce what Edward Said theorized as Orientalism [22], insofar as marginalized groups become objects of representation rather than epistemic agents. Even inclusive efforts to incorporate "diverse values" may unintentionally transform situated and evolving social relations into extractable data representations.

To move beyond this limitation, we draw on intersectional theory [23], which emphasizes that social positions are internally differentiated and structurally produced. Class, gender, labor relations, linguistic marginalization, and technological exposure jointly shape fundamentally different relations to AI systems, even within the same nominal group [24]. For example, the lived experience of a highly paid urban technology worker and a precarious content moderator may diverge sharply within the same national context.

We therefore suggest that alignment discourse would benefit from shifting attention away from identity-based aggregation toward relational positioning within sociotechnical systems.

4.3 Toward Relational Pluriversal Alignment

Building on this critique, we advance a relational interpretation of pluriversal alignment grounded in algorithmic coloniality. Rather than treating alignment as the aggregation of pre-given preferences, it can be understood as a normative reorientation of how epistemic authority is distributed and recognized within sociotechnical systems.

This perspective suggests three normative commitments.

Commitment 1: Affectedness over Representation. Epistemic authority should be more attentive to degrees of systemic exposure to AI-related harms and externalities. Those who bear disproportionate risks—such as content moderators, marginalized language communities, or populations subject to automated misclassification—should be meaningfully included in shaping and contesting system outcomes.

Commitment 2: Relational Position over Essentialist Identity. Alignment should move beyond static demographic categories and attend instead to relational positions within sociotechnical infrastructures, including labor conditions, platform dependency, and structural vulnerability. This helps avoid reifying cultural identities while foregrounding the production of inequality through technical systems.

Commitment 3: Solidarity over Preference Aggregation. Rather than aggregating preferences across abstract groups, alignment can be understood as being informed by forms of translocal solidarity grounded in shared exposure to systemic risk. Such solidarity is not based on shared identity, but on overlapping relations of exploitation, extraction, and epistemic marginalization.

Taken together, these commitments reframe alignment not as a fixed technical solution but as a normative orientation toward sociotechnical systems under conditions of structural asymmetry. In this sense, pluriversal alignment functions less as a design blueprint and more as a critical lens for rethinking what it means for AI systems to be aligned at all.

5 Conclusion

This paper has argued that AI alignment cannot be adequately understood as a neutral problem of aggregating human preferences. Instead, it must be situated within the historical and structural conditions of coloniality embedded in contemporary socio-technical systems.

We have shown that AI infrastructures are not neutral optimization systems, but are shaped by and participate in global asymmetries of labor, knowledge production, and risk distribution. These asymmetries affect not only system

outcomes, but also the conditions under which preferences and values are formed and rendered computable.

From this diagnostic perspective, we have proposed a relational reinterpretation of alignment that shifts attention from universal preference aggregation toward epistemic authority and structural exposure within AI systems. This reframing emphasizes how socio-technical systems distribute voice, visibility, and harm across differently situated populations.

More broadly, this perspective suggests that AI alignment is not merely a technical problem, but a site where competing normative orders of knowledge and authority are negotiated and made visible.

References

1. Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4299–4307 (2017)
2. Hadfield-Menell, D., Russell, S., Abbeel, P., Dragan, A.: Cooperative inverse reinforcement learning. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 3909–3917 (2016)
3. Russell, S.: Human compatible: AI and the problem of control. Penguin Uk (2019)
4. Wang, J., Xue, S., Li, J., Wang, X.: Diverse Human Value Alignment for Large Language Models via Ethical Reasoning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, no. 3, pp. 2637–2648 (2025)
5. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and Abstraction in Sociotechnical Systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), pp. 59–68. ACM, New York (2019)
6. Noble, S.U.: Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, New York (2018)
7. Couldry, N., Mejias, U.A.: The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism. (2020)
8. Mohamed, S., Png, M.T., Isaac, W.: Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology **33**(4), 659–684 (2020)
9. Mignolo, W.D.: Delinking: The Rhetoric of Modernity, the Logic of Coloniality and the Grammar of De-Coloniality. Cultural Studies **21**(2–3), 449–514 (2007)
10. Quijano, A.: Coloniality of Power, Eurocentrism, and Latin America. Nepantla: Views from South **1**(3), 533–580 (2000)
11. Prebisch, R.: The Economic Development of Latin America and Its Principal Problems. United Nations, New York (1962/2021)
12. Benjamin, R.: Race after Technology: Abolitionist Tools for the New Jim Code. Polity, Cambridge (2019)
13. Srnicek, N.: Platform Capitalism. John Wiley & Sons (2017)
14. Casilli, A.A.: Digital Labor Studies Go Global: Toward a Digital Decolonial Turn. International Journal of Communication **11**, 3934–3954 (2017)
15. Gray, M.L., Suri, S.: Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Harper Business, New York (2019)

16. Hadwick, D., Lan, S.: Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany. World Tax Journal, 609–640 (2021)
17. Perrigo, B.: OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic. Time Magazine, 18, 2023.
18. Escobar, A.: Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds. Duke University Press, Durham (2018)
19. Baum, K., Slavkovik, M.: Aggregation Problems in Machine Ethics and AI Alignment. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, no. 1, pp. 355–366 (2025)
20. Leyva, M., et al.: Pluriversal Alignment: Paraconsistency, Latin American Logic, and the Decolonial Critique of Artificial Intelligence (2026)
21. Radiya-Dixit, E., Christin, A.: Same Stereotypes, Different Term? Understanding the "Global South" in AI Ethics. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, no. 3, pp. 2081–2093 (2025)
22. Said, E.W.: Orientalism. Pantheon Books, New York (1978)
23. Crenshaw, K.: Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. In: Bartlett, K.T., Kennedy, R. (eds.) Feminist Legal Theories, pp. 23–51. Routledge, London (2013)
24. Mohanty, C.T.: Under Western Eyes: Feminist Scholarship and Colonial Discourses. Feminist Review **30**, 61–88 (1988)
25. Birhane, A.: Algorithmic Colonization of Africa. SCRIPTed **17**, 389–409 (2020)
26. Crawford, K.: The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)
27. Posada, J.: The Coloniality of Data Work in Latin America. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 277–278 (2021)
28. Jobin, A., Ienca, M., Vayena, E.: The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence **1**(9), 389–399 (2019)
29. Mitchell, M., et al.: Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229 (2019)
30. Harvey, D.: The New Imperialism. Oxford University Press (2003)
31. Edwards, P.N.: The Closed World: Computers and the Politics of Discourse in Cold War America. MIT Press, Cambridge (1996)
32. Ensmenger, N.: The Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise. MIT Press, Cambridge (2012)
33. Winner, L.: Do Artifacts Have Politics? In: Johnson, D.G., Nissenbaum, H. (eds.) Computer Ethics, pp. 177–192. Routledge, London (2017)