Universal Alignment Fails in Global Classrooms: Cross-Cultural Blind Spots in EdTech AI

Zijin Wu [1] **David Scott Lewis** [1]

Abstract

The global expansion of educational AI often falls into a "portability trap," where dominant EdTech systems aligned to WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations impose culturally specific assumptions on diverse global classrooms. We argue that these cross-cultural failures are fundamentally alignment problems: standard pipelines like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) collapse legitimate educational disagreements over pedagogy, privacy, and language into a single, culturally rigid reward target. To bridge theory and practice, we map documented classroom harms—linguistic marginalization, biometric proctoring bias, pedagogical mismatch, and socio-emotional neglect—directly to specific ML failure modes, including distributional collapse and reward misspecification. In response, we propose a four-pillar socio-technical roadmap for pluralistic EdTech: modular objective design, sovereign data infrastructures, localized impact verification, and human mediation. Ultimately, responsible educational AI requires a universal rights floor paired with pluralistic systems that can be explicitly steered toward legitimate local values, rather than assimilating them into an algorithmic monoculture.

1. Introduction

The rapid integration of artificial intelligence (AI) into education is often presented as a global public good: a way to personalize instruction, expand access, automate assessment, and reduce teacher shortages. Yet educational AI is not culturally neutral. It is shaped by the data, objectives, institutions, and values of its designers. In practice, the EdTech ecosystem remains concentrated in the Global North, and many dominant models are trained, aligned, and evaluated on data that overrepresent Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations and pedagogical norms (Porayska-Pomsta et al., 2023; Mouta et al., 2024; Samuel et al., 2023).

When these systems are deployed across diverse educational settings, they often fall into the portability trap: systems and governance assumptions developed in one context are transferred into another without sufficient adaptation to local languages, pedagogies, infrastructures, or ethical priorities (Schiff, 2022; Hulus, 2026; Arif, 2025). The effects are concrete: Generative AI produces Western-centric examples in non-Western classrooms; speech systems misrecognize non-standard dialects and regional languages; and automated proctoring disproportionately misclassifies darker-skinned students and students in low-resource environments (Nyaaba et al., 2024; Koenecke et al., 2020; Buolamwini & Gebru, 2018; Yoder-Himes et al., 2022; R et al., 2025). Recent field surveys confirm these harms, with students frequently reporting linguistic self-censorship and cultural alienation when forced to adapt to foreign algorithmic norms (R et al., 2025).

This position paper does not introduce a new benchmark or model; rather, it argues that current alignment pipelines are structurally ill-suited to global educational deployment and that pluralistic alignment is the correct technical and governance response. In this paper, a cross-cultural ethical blind spot is a model behavior that appears acceptable under one educational value regime but fails under another; pluralistic alignment means preserving multiple legitimate educational preference profiles rather than collapsing them into one reward target; and the portability trap refers to exporting models, benchmarks, and governance assumptions across contexts without adequate adaptation.

We argue that these are not only fairness failures but alignment failures. Current alignment methods such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) usually aggregate feedback into a single preference model, treating disagreement over pedagogy, surveillance, and assessment as statistical noise rather than legitimate value pluralism. More concretely, the

[1]AI Executive Consulting, Zaragoza, Aragon, Spain. Correspondence to: Zijin Wu <ryan.wu@aiexecutiveconsulting.com>.

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

alignment problem here is the choice of whose educational preferences become the model's default objective when heterogeneous feedback is aggregated into a single policy. Pluralistic alignment research shows that standard alignment can reduce distributional pluralism by homogenizing outputs (Sorensen et al., 2024). Social choice theory sharpens the point: under conflicting preferences, no single aggregation rule can satisfy all desirable fairness conditions, so a universally aligned model is conceptually unstable rather than merely technically difficult (Mishra, 2023). Value pluralism therefore matters because educational conflict is not something to be averaged away; it is part of the domain itself (Kasirzadeh, 2024).

This challenge is acute in education because educational AI helps define which forms of language, conduct, and achievement are recognized as legitimate. Core ethical concerns include pedagogy, emotional well-being, accountability, teacher autonomy, and cultural fit, not only privacy and bias (Porayska-Pomsta et al., 2023; Mouta et al., 2024; Khan, 2023). These concerns also vary by region. While Western frameworks emphasize privacy and learner autonomy, some East Asian contexts prioritize collective harmony, and many African and South Asian contexts foreground equity, multilingual inclusion, and infrastructural feasibility (Hulus, 2026; Arif, 2025; Samuel et al., 2023; Mumtaz et al., 2025; Eryilmaz, 2026). Because authoritative AI systems can also shape user beliefs and judgments, these blind spots are not only representational but pedagogical (Kidd & Birhane, 2023).

Commercial incentives deepen the problem. Vendors are rewarded for deploying single scalable models, mirroring patterns of data colonialism, where behavioral and linguistic data is extracted often from the Global South to improve foreign-controlled systems (Couldry & Mejias, 2019; Kohnke & Foung, 2024). In education, this can become digital neocolonialism, where students must adapt to foreign algorithmic norms while their data strengthen the systems that marginalize them (Nyaaba et al., 2024; R et al., 2025).

We therefore argue that a single universally aligned educational AI is a mathematical, sociotechnical, and political illusion. Global EdTech instead requires pluralistic alignment. Following Sorensen et al. (2024), we use this term for alignment methods that preserve value diversity rather than collapse it into one reward. Crucially, this is distinct from mere localization. While localization risks replacing a global monoculture with a regional one—for example, forcing diverse Indigenous learners to conform to a single dominant national standard (R et al., 2025)—pluralistic alignment preserves legitimate subgroup variations (*distributional pluralism*), surfaces multiple defensible perspectives (*Overton pluralism*), and allows local educators to dynamically adapt model behavior to navigate competing values (*steerable pluralism*).

Unlike prior AIED ethics reviews that mainly catalogue harms, this paper reframes cross-cultural EdTech failures as an alignment problem (Mouta et al., 2024; Hulus, 2026; Porayska-Pomsta et al., 2023). Specifically, this paper makes three core contributions: (1) we map major cross-cultural harms in EdTech to failure modes in the ML pipeline, (2) we integrate decolonial critique and thick models of value with pluralistic alignment, and (3) we outline a socio-technical roadmap spanning modular alignment architectures, localized governance, Indigenous data sovereignty, and culturally adapted impact assessment.

Our core claim is simple: if educational AI is to serve global classrooms rather than export a single worldview into them, alignment must become explicitly pluralistic.

2. Background: The Illusion of Universal Alignment

The assumption that educational values—teaching, assessment, privacy, and support—can be aggregated into a single optimization target and deployed globally is weak on both technical and sociological grounds. Technically, RLHF and DPO typically compress human judgments into a single reward model. Sociologically, the data and institutions through which those judgments are collected are shaped by unequal power, language dominance, and platform concentration. In education, these problems reinforce each other (Porayska-Pomsta et al., 2023; Mouta et al., 2024).

2.1. Social Choice, Preference Heterogeneity, and the Limits of Consensus

Standard alignment pipelines treat human feedback as if it revealed a coherent social welfare function, and the system is optimized toward one behavioral target. That architecture is efficient but makes the fragile assumption that disagreement can be reconciled into consensus. What counts as a "good" response in education depends on contested pedagogical norms: whether AI should encourage deference or critique, whether monitoring is protective or invasive, or whether linguistic conformity should be rewarded.

Social choice theory explains why such disagreement cannot be cleanly collapsed into one objective. Arrow's theorem shows that with at least three alternatives, no aggregation rule satisfies a standard set of fairness conditions for all preference profiles (Arrow, 1951). Recent alignment work applies this directly to AI: when groups rank outputs differently, any single reward model must privilege some preferences over others (Mishra, 2023). Monolithic alignment therefore encodes a particular compromise rule instead of a universal agreement.

Pluralistic alignment work makes this more concrete. Sorensen et al. (2024) argue that standard alignment can reduce distributional pluralism by homogenizing outputs that would otherwise reflect heterogeneous viewpoints. Kasirzadeh (2024) likewise argues that value conflicts should not be treated as annotation noise to be averaged away. Recent work on DPO with unobserved preference heterogeneity reinforces the finding that simple aggregated comparisons can fail to recover legitimate subgroup-specific preferences in the first place (Chidambaram et al., 2025). In education, a single aligned model therefore encodes a dominant educational settlement while obscuring the trade-offs it imposes.

2.2. Data Colonialism, Annotation Politics, and Epistemic Injustice

The impossibility of universal consensus is intensified by the unequal political economy of data. Couldry & Mejias (2019) describe data colonialism as a form of extraction in which human life is appropriated through data capture and converted into value for distant platform owners. In EdTech, students are not only users of AI systems but also continuous sources of behavioral data. Kohnke & Foung (2024) show that educational platforms often normalize this extraction while offering limited local control or meaningful consent.

Crucially, such extraction does not yield authentic representation. Pretraining corpora and reward models still overrepresent English-dominant, Western contexts. Even when Global South data is captured, alignment pipelines resolve disagreement via majority vote or dominant-language standards, recasting alternative epistemologies as statistical error rather than valid pluralism. This causes epistemic injustice: learners are denied recognition as credible knowers because the system enforces a narrow cultural grammar of legitimacy (Fricker, 2007; Kay et al., 2024; Mollema, 2025; Porayska-Pomsta et al., 2023).

Recent work illustrates this paradox: extraction under WEIRD-aligned models forces user assimilation. Nyaaba et al. (2024) show widely used GenAI tools return Western-centric content while underrepresenting local histories. Forced to navigate these systems, students adapt to the machine. In Indian higher education, only 35.41% of students felt AI reflected their reality, and 85% reported rephrasing their language to sound more "standard" to avoid misrecognition (R et al., 2025). Thus, extracted data often reflects linguistic self-censorship rather than local culture, inadvertently strengthening the models that marginalize them. These are the downstream effects of training distributions and data ownership patterns favoring dominant educational worlds.

2.3. Regional Value Variation Beyond the West/Non-West Binary

Furthermore, cross-cultural ethics cannot be reduced to a simple West/non-West binary. Cross-cultural research shows a more complex landscape.

Western and North American frameworks often emphasize privacy, learner autonomy, and anti-bias protections. Some East Asian contexts place greater weight on collective harmony, social obligation, and efficiency. Many African and South Asian contexts foreground equity, access, multilingual inclusion, and infrastructural feasibility (Hulus, 2026; Arif, 2025; Samuel et al., 2023). These are not fixed civilizational essences; they show that educational values vary across multiple dimensions and institutional histories.

Empirical studies support this view. Mumtaz et al. (2025) find significant cross-cultural differences in how students evaluate the ethical use of AI tools in higher education. Eryilmaz (2026) similarly reports that cultural dimensions such as uncertainty avoidance and long-term orientation influence whether AI is viewed mainly as an opportunity or a risk. Ethical variation therefore cannot be handled through translation alone. It requires alignment methods and governance processes that are sensitive to pedagogy, emotional norms, language ecologies, and material infrastructure.

In the remainder of the paper, these background constraints reappear as four concrete ML failure modes: distributional collapse in language systems, normative overfitting in proctoring, reward misspecification in pedagogical optimization, and objective-function blindness to socio-emotional harms.

3. Core Explorations: Mapping Cultural Harms to ML Failures

To move beyond generic discussions of "AI bias," we must connect classroom harms to specific failures in the ML pipeline. As illustrated in Figure 1, cross-cultural harms in global EdTech arise from a "portability trap" pipeline where Global North-centered inputs translate into concrete institutional harms via specific ML failures. Four recurring pathologies are especially salient: distributional collapse in language technologies, intersectional representation deficits in biometric systems, reward misspecification in pedagogical optimization, and objective-function blindness to socio-emotional safety (Porayska-Pomsta et al., 2023; Sorensen et al., 2024).

3.1. Linguistic Marginalization vs. Distributional Collapse

The core mistake in current language-alignment pipelines is treating culturally legitimate linguistic variation as low-probability noise. Educational AI increasingly relies on

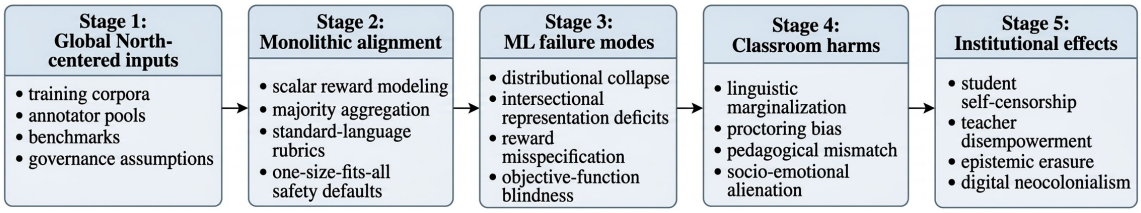

Figure 1. The Portability Trap Pipeline

natural language technologies: automated essay scoring, tutoring dialogue, speech recognition, translation, and writing feedback. Yet pretraining corpora for language and speech systems overrepresent standard, professional, and often Western forms of English, pushing dialectal variation, Indigenous languages, and regional rhetorical styles to the tails of the distribution (Porayska-Pomsta et al., 2023; Mayfield et al., 2019). The result is linguistic marginalization, where culturally situated language is treated as less intelligible and less legitimate.

Alignment and evaluation can worsen the problem. Reward models and rubrics often encode local norms of "clarity" and "fluency" as neutral quality markers when they are not. Preference heterogeneity research shows that aggregated comparisons can treat diverse rhetorical tones as errors rather than hidden context, especially when annotators evaluate tone, rhetoric, and communicative purpose under different assumptions (Chidambaram et al., 2025; Bahlous-Boldi et al., 2024).

Empirical evidence shows the scale of the problem. Automated speech recognition systems have substantially higher error rates for Black speakers than for white speakers (Koenecke et al., 2020). In Indian higher education, R et al. (2025) found that 53.10% of students experienced accent- or mother-tongue-related AI misrecognition, while tribal and Indigenous speakers faced a 58.3% "very frequent" misrecognition rate, compared with 10% for English speakers. Automated grading and writing support show that essays using non-standard English, culturally specific examples, or alternative rhetorical structures are often penalized even when the reasoning is strong (Khan, 2023; Mayfield et al., 2019). In practice, this erasure pressures students to sound "more standard" to the machine rather than more culturally authentic.

3.2. Biometric Proctoring vs. Intersectional Representation

Proctoring harms arise because the system learns a narrow normative model of visibility and legitimacy, then mistakes contextual variation for misconduct. These systems use computer vision to detect faces, track gaze, classify movement, and flag anomalous behavior during high-stakes assessments. Their harms stem from two interacting failures. The first problem is intersectional representation. Buolamwini & Gebru (2018) showed that commercial facial analysis systems performed worse on darker-skinned women because benchmark datasets overrepresented lighter-skinned subjects. Yoder-Himes et al. (2022) extend this concern directly to educational proctoring, documenting racial, skin-tone, and sex disparities. This is a standard out-of-distribution (OOD) generalization failure.

The second problem, however, is a fundamental alignment failure: normative anomaly detection. While ML literature often treats cross-cultural computer vision failures merely as domain shift problems requiring more training data, defining what constitutes an "anomaly" is effectively a reward specification problem. Proctoring models optimize for conformity to a fixed behavioral policy (e.g., constant screen gaze, isolated quiet rooms). Consequently, when encountering culturally normative gaze aversion or shared low-resource households, simply deploying a "more robust" model does not mitigate the harm; it merely penalizes local realities more accurately (R et al., 2025; Arif, 2025). Pluralistic alignment is uniquely required here because the behavioral objective itself—what constitutes legitimate test-taking—is culturally contested, not universally fixed.

These harms are therefore not only fairness failures but failures of context sensitivity and process fairness (Porayska-Pomsta et al., 2023). In India, 60.16% of surveyed students reported severe discomfort with AI proctoring, and marginalized students described disproportionate flagging and stress under low-light and space-constrained conditions (R et al., 2025). Biometric proctoring thus embeds a narrow normative model of visibility, composure, and legitimacy that breaks under cross-cultural domain shift.

3.3. Pedagogical Misalignment vs. Reward Misspecification

Pedagogical misalignment is when a culturally plural task is optimized toward one scalar objective. If language systems misrecognize expression and proctoring systems misread context, AI learning platforms often fail one layer deeper: they optimize for the wrong educational objective. Many EdTech systems are built around easy-to-measure proxies such as test-score improvement, click-through rates, or engagement. These metrics are poor stand-ins for the full

range of educational goals. In ML terms, this is reward misspecification.

Educational success is not a single scalar quantity. Different cultural traditions weigh mastery, collaboration, teacher authority, critical debate, and moral development differently (Samuel et al., 2023). Pluralistic alignment demonstrates that optimizing a multi-objective task as a single scalar target invariably privileges a dominant behavioral model (often Western, individualized, self-directed learning) while undermining relational or collective learning environments (Vamplew et al., 2024; Farheen et al., 2025).

Pluralistic alignment work indicates that educational AI is not balancing one reward but many potentially competing ones (e.g., accuracy, autonomy, teacher authority, and collaboration) (Vamplew et al., 2024; Li et al., 2025; Xiong & Singh, 2025). From this perspective, current EdTech systems are often optimizing an impoverished objective that excludes central educational values from the start.

Existing AIED ethics research supports this diagnosis. Porayska-Pomsta et al. (2023) warn that AI systems often embed inflexible pedagogies. Samuel et al. (2023) argue that culturally adaptive educational AI must account for differences in autonomy, knowledge-sharing, and classroom expectations. Arif (2025) similarly shows that teacher-centered educational cultures may experience individualized AI systems as disruptive rather than empowering. Farheen et al. (2025) add that overreliance on quantitative indicators sidelines creativity, belonging, and critical reflection. Pedagogical misalignment is therefore the downstream result of optimizing narrow proxies for a culturally heterogeneous task.

3.4. Socio-Emotional Neglect vs. Objective-Function Blindness

Finally, socio-emotional neglect follows when educational appropriateness is treated as correctness plus engagement, rather than as a culturally mediated relational good. Educational AI shapes how students experience recognition, encouragement, embarrassment, and trust. Yet socio-emotional safety is weakly represented in objective functions: Mouta et al. (2024) find that only a small fraction of the AIED ethics literature addresses emotional management, despite the obvious effects of AI-mediated learning on anxiety, confidence, and belonging.

This neglect is especially consequential for marginalized learners, who are exposed to repeated algorithmic misrecognition. Khan (2023) reports that students subject to biased AI systems commonly experience frustration, disengagement, and feelings of inadequacy. R et al. (2025) document similar effects in Indian higher education, where linguistic correction and cultural misrecognition generate alienation and performance barriers. These harms can worsen when educational AI adopts affective styles that do not travel well across cultures. A chatbot calibrated for direct, therapeutic, or highly informal emotional support may violate local norms around respect, hierarchy, emotional restraint, or the role of teachers and peers (Hulus, 2026; Samuel et al., 2023; Pekrun et al., 2023).

From an ML perspective, this is a specific pathology we term *objective-function blindness*, which occurs when critical dimensions of human well-being—such as preserving student dignity or avoiding shame—are entirely absent from the model's loss landscape. Because these affective dimensions are unmeasured, the model remains blind to socio-emotional collateral damage. Emerging work on steerable and personalized pluralistic alignment suggests that some of these dimensions could be modeled explicitly by conditioning on user or community preferences, but current EdTech systems rarely do so in a principled way (Feng et al., 2025).

Taken together, these four cases show that cross-cultural harms in EdTech arise from core ML choices about representation, normalization, reward design, and evaluation. Post hoc fairness patches are therefore insufficient.

4. The Governance & Technical Void

Cross-cultural harms in EdTech are sustained by a structural vacuum in technical interpretability and institutional governance. When deployed globally, these systems operate in a void of algorithmic opacity, mismatched regulatory exports, and exclusionary top-down procurement. This creates a second-order alignment failure: affected communities lack the documentation, appeal channels, and authority to contest the optimized values (Porayska-Pomsta et al., 2023; Mouta et al., 2024; Farheen et al., 2025).

4.1. The Black Box in the Classroom

The black-box nature of deep learning creates an educational accountability crisis. Classrooms require feedback, explanation, and contestability, yet the limited interpretability of language and vision models prevents educators and students from understanding automated decisions. When penalized by automated scorers or biometric proctors, neither students nor teachers can meaningfully inspect reasoning or challenge judgments. Beyond poor interpretability, closed models suffer from weak auditability in practice, offering limited documentation, no meaningful appeal paths, and restricted access for independent review (Farheen et al., 2025; Porayska-Pomsta et al., 2023; Mouta et al., 2024). Accountability is thus displaced from local educators toward opaque, transnational vendors.

4.2. The Illusion of Universal Regulation

To address these accountability gaps, policymakers often treat Western frameworks—like the EU AI Act or GDPR—as universal gold standards (European Union, 2024). While providing important rights-based baselines, exporting them unchanged reproduces the portability trap. Such frameworks assume European regulatory capacities and prioritize individual privacy and autonomy, which may misalign with African or South Asian contexts emphasizing communal equity, multilingual inclusion, and infrastructural access (Hulus, 2026; Mouta et al., 2024; Zeide & Nissenbaum, 2018). Moreover, these frameworks assume auditing capacity, institutional literacy, and enforcement resources typically scarce in lower-income school systems. Without localized human-rights or educational impact assessments, universal regulation devolves into procedural compliance lacking substantive protection (Hulus, 2026; Ceravolo et al., 2025).

4.3. Top-Down Procurement and the Lack of Participatory AI

Without culturally enforceable regulation, EdTech governance defaults to corporate procurement driven by transnational vendors and administrators, bypassing the communities living with the system's effects. Practically, ethical "alignment" is set upstream in training pipelines and product decisions rather than local contexts of use. This top-down model excludes marginalized groups from design, evaluation, and deployment. Treating Global South students primarily as data subjects strips communities of epistemic agency and reinforces digital neocolonialism (Kohnke & Foung, 2024; Nyaaba et al., 2024).

Moving beyond mere consultation requires upstream participatory governance: community-defined value elicitation, localized impact assessment, and shared oversight before deployment (Mayer, 2025; Ter-Minassian, 2025). This means shifting from vendor-defined to co-governed procurement: contracts must require independent local review bodies, public disclosure of failure modes, documented appeals, and vendor-funded remediation for high-risk deployments.

5. Roadmap: Operationalizing Pluralistic EdTech

Diagnosing algorithmic monoculture is insufficient without implementable alternatives. If the cross-cultural harms in Section 3 arise from a monolithic ML pipeline (Figure 1), and are sustained by the institutional gaps in Section 4, then a viable roadmap must redesign both layers simultaneously. We therefore propose a four-pillar socio-technical agenda, where each pillar systematically dismantles a specific component of the portability trap:

- Modular objective design replaces monolithic reward pipeline (Stage 2)
- Sovereign data infrastructures replace extractive, Global North-centered inputs (Stage 1)
- Localized impact verification fills the regulatory void (Section 4.2)
- Human mediation directly counters the black-box accountability crisis (Section 4.1)

As outlined in Table 1, these pillars form a matrix, providing specific, actionable mitigations across all four classroom harm domains. The goal is not cosmetic localization after deployment, but a shift toward systems that can represent, preserve, and govern legitimate educational disagreement (Sorensen et al., 2024; Hulus, 2026). Each pillar also implies a different responsible actor: ML developers redesign objectives, institutions govern data and procurement, regulators enforce impact assessments, and educators mediate and contest outputs in practice.

5.1. From Monolithic Rewards to Modular Pluralism

Addressing the monolithic alignment of Stage 2 (Figure 1), the standard RLHF paradigm collapses diverse pedagogical values across communities into a single scalar reward to represent a homogeneous "standard" user (Vamplew et al., 2024; Li et al., 2025; Xiong & Singh, 2025). A useful system must recognize that diverse dimensions (e.g., cultural relevance, collaboration, dignity, accessibility) are weighted differently across classrooms and communities. This is precisely the setting targeted by recent work in multi-objective reinforcement learning, multi-objective alignment for LLMs, and multi-group RLHF, all of which treat disagreement as structural rather than exceptional (Vamplew et al., 2024; Li et al., 2025; Xiong & Singh, 2025).

Operationally, developers should replace one global reward model with modular or steerable architectures. Approaches like multi-objective preference optimization (MODPO) model educational objectives separately rather than collapsing them at training time (Zhou et al., 2024). Alternatively, steerable pluralism conditions behavior at inference time on community-specified value profiles (Feng et al., 2025). For high-stakes tasks like tutoring, developers can deploy multi-agent committee architectures where distinct agents evaluate outputs across different criteria (e.g., factual correctness versus emotional tone). Crucially, aggregation must expose trade-offs rather than erase minority perspectives through simple majority voting. In practice, the output should be either a context-sensitive recommendation or a transparently plural set of defensible responses, not a falsely universal answer. Consequently, models must be evaluated on distributional and steerable pluralism across subgroups rather

Table 1. Four-Pillar Roadmap for Pluralistic EdTech

Failure Modes	Intervention Pillars			
	Pluralistic Objective Design	**Sovereign Data Infrastructures**	**Localized Impact Verification**	**Human Mediation & Contestation**
Linguistic Marginalization	Multi-objective / steerable language evaluation	Multilingual local datasets / federated adaptation	Disaggregated performance testing by language community	Community-defined value elicitation
Biometric Proctoring Bias	Context-sensitive recommendation models	Differential privacy and secure aggregation	Subgroup audits / low-light and shared-device stress tests	Accessible appeal channels for flagged behavior
Pedagogical Misalignment	Multi-objective preference optimization (MODPO)	Community-curated pedagogical datasets	Context sensitivity audits against local pedagogical values	Teacher override and local review
Socio-Emotional Neglect	Affect-sensitive steerable response policies	Privacy-preserving local adaptation	Post-deployment monitoring for socio-emotional harms	Participatory workshops and community-facing support

than just average benchmark accuracy.

5.2. Implementing Indigenous Data Sovereignty

To replace the Global North-centered inputs that drive the portability trap (Kohnke & Foung, 2024; Nyaaba et al., 2024), control over data collection, curation, and adaptation must shift to local communities. This aligns with Indigenous Data Sovereignty and the CARE principles (Collective benefit, Authority to control, Responsibility, Ethics) (Kukutai & Taylor, 2016; Russo Carroll et al., 2021). While originating in settler-colonial contexts (e.g., Australia, Canada, the US), CARE's core mechanisms of anti-extraction and community-governed data stewardship offer a blueprint for Global South educational settings resisting data extraction.

Technically, this requires distributed alignment methods like federated learning, allowing local institutions to fine-tune adapters on representative multilingual data while keeping raw student data in-country. Recent work on federated learning and adaptive preference pluralistic alignment suggests that distributed training can preserve local preference variation more effectively when different communities hold distinct value priorities (Srewa et al., 2025; 2026). However, data localization alone is not enough. Federated systems must incorporate *differential privacy*—mathematical guarantees that individual student data cannot be reverse-engineered or memorized during federated learning—and *participatory governance*—frameworks where marginalized educators and students hold decision-making power over data collection and labeling. Without these dual protections, local data collection risks reproducing regional power hierarchies and continues to make minority groups underrepresented (Yang & Al-Masri, 2025).

5.3. Localized Fundamental Rights Impact Assessments

To fill the regulatory void left by mismatched universal frameworks (Section 4.2), educational deployment requires localized and continuous verification. We therefore propose that ministries, universities, and school systems require culturally adapted Fundamental Rights Impact Assessments (FRIAs) before procurement and after deployment (Ceravolo et al., 2025; Mantelero, 2024; Ullstein et al., 2025).

For any system used in high-stakes grading, biometric monitoring, or deployments above a defined student-user threshold, the vendor should fund the FRIA, an independent evaluator should conduct it, and a short public report should disclose subgroup error patterns, contestation procedures, and required remediation steps. Rather than paperwork checklists, these lifecycle evaluations must test subgroup performance, stress-test infrastructure under realistic conditions (e.g., low bandwidth, shared devices), and assess context sensitivity, contestability, and post-deployment drift. If a system fails these localized tests, applications like biometric proctoring must be restricted in those settings (R et al., 2025; Mouta et al., 2024; European Union, 2024) (see Table 2 in the Appendix for a detailed operational checklist).

5.4. Digital Literacy as Cultural Mediation

To counter the black-box opacity of classrooms (Section 4.1) and top-down procurement (Section 4.3), technical reforms must be paired with human mediation. Digital literacy must be redefined as cultural mediation, equipping teachers and students to interpret AI through local norms and educational purposes rather than treating outputs as culturally neutral (Hulus, 2026; Samuel et al., 2023).

In practice, beyond prompt engineering, educators require training and institutional authority to inspect model prove-

nance, override misaligned outputs, and feed local critiques back into procurement pipelines. Students similarly need instruction that AI outputs are situated, contestable, and shaped by training data rather than objective truth (Kidd & Birhane, 2023). In low-resource settings, where infrastructural constraints limit appeal opportunities, local review committees and participatory workshops are essential to convert users from passive data subjects into co-governors of educational AI (Judijanto, 2025; Mouta et al., 2024).

6. Limitations and Ethical Tensions

While addressing algorithmic monoculture, pluralistic alignment introduces the paradox of tolerance: pluralism does not imply moral relativism. Adapting AI to local norms risks inadvertently codifying historical prejudices, such as patriarchal or caste-based exclusions embedded within local institutions. Culturally adaptive AI therefore requires explicit boundary conditions ensuring student safety, non-discrimination, and basic rights. In practice, this means pairing local adaptation with rights-based impact assessment, explicitly acknowledging the paradox that defining a "universal rights floor" (often rooted in UN/Western legal frameworks) risks repeating the very top-down exportation we critique.

Navigating this requires treating fundamental rights as a dynamically negotiated baseline rather than an absolute, externally imposed edict (Ceravolo et al., 2025; Hulus, 2026; Kasirzadeh, 2024). Operationally, this negotiation is executed through the Localized FRIA process, where independent committees—structurally prioritizing marginalized learner voices to prevent dominant-group capture—translate abstract rights into context-specific thresholds. Anchored by co-governed procurement contracts, these committees exercise the institutional authority to mandate vendor remediation or veto high-risk deployments if these community-defined baselines are violated.

A second tension concerns scalability. The economic logic of current foundation models favors centralized APIs, whereas pluralistic EdTech demands modular, federated, and locally adapted systems. These approaches are more culturally responsive, but they are also more computationally demanding and operationally complex. For many educational systems in the Global South, where local technical capacity are already constrained, the costs of maintaining bespoke pluralistic systems may be prohibitive (Judijanto, 2025; Hulus, 2026). Bridging this gap requires lightweight ML techniques like Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA). By allowing base models to be locally steered via lightweight adapters, communities can maintain meaningful local control without the prohibitive costs of full-parameter training.

Third, "the community" is never a single actor. Local participation risks capture by dominant language groups, higher-caste actors, or politically organized stakeholders whose preferences exclude the most vulnerable learners. Pluralistic alignment must therefore incorporate subgroup protections and transparent conflict-resolution procedures when local values themselves are contested (Mayer, 2025; Ter-Minassian, 2025; Sloane et al., 2022).

A further risk is false decentralization. Open-weight or locally fine-tuned systems can still inherit Western reward shaping, inaccessible inference infrastructure, and centralized procurement dependence. True pluralism requires local authority over objectives and deployment, not merely open access.

Finally, current evidence of cross-cultural AI harms in literature relies heavily on cross-sectional surveys and qualitative research. While indispensable, the field lacks robust pluralism-specific benchmarks, cross-lingual stress tests, and longitudinal post-deployment evaluations. Future work should combine participatory methods with controlled subgroup benchmarking, culturally aware evaluation suites, and ongoing monitoring of steerability, context sensitivity, and downstream harms in real classrooms (Mouta et al., 2024; Sorensen et al., 2024).

7. Conclusion

The global expansion of EdTech AI is often celebrated for expanding access, yet deploying generic, Western-aligned models into global classrooms risks enforcing algorithmic monoculture. When systems penalize linguistic diversity, normalize biometric surveillance, and privilege Western pedagogy, they marginalize local epistemologies and deepen digital neocolonialism (R et al., 2025; Nyaaba et al., 2024). Technological scaling must not supersede cultural safety.

The remedy is a layered alignment model: a universal rights floor—acknowledged as a negotiated baseline rather than a rigid Western export—paired with pluralistic systems steerable toward local realities (Sorensen et al., 2024; Hulus, 2026; Ceravolo et al., 2025). Modular multi-agent architectures, localized impact assessments, Indigenous data sovereignty, and community oversight are not isolated reforms; together, they establish the socio-technical conditions for culturally sustaining EdTech.

Ultimately, no single benchmark, data regime, or market actor should silently define what constitutes a "good" education globally. Pluralistic alignment must evolve from abstract critique into an auditable design principle and governance standard, ensuring AI serves diverse classrooms rather than assimilating them.

References

Arif, S. Cross-cultural perspectives on ai in education: Case studies from global classrooms. *Artificial Intelligence in Education Journal*, 2025.

Arrow, K. J. *Social Choice and Individual Values*. John Wiley & Sons, 1951.

Bahlous-Boldi, R., Ding, L., Spector, L., and Niekum, S. Pareto-optimal learning from preferences with hidden context. arXiv preprint, 2024.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pp. 77–91. PMLR, 2018.

Ceravolo, P. et al. Hh4ai: A methodological framework for ai human rights impact assessment. arXiv preprint, 2025. Full list of 12 authors available on arXiv.

Chidambaram, K., Seetharaman, K. V., and Syrgkanis, V. Direct preference optimization with unobserved preference heterogeneity. arXiv preprint, 2025.

Couldry, N. and Mejias, U. A. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349, 2019. doi: 10.1177/1527476418796632.

Eryilmaz, M. A cross-cultural examination of ethical issues in ai development. *PeerJ Computer Science*, 12:e3504, 2026. doi: 10.7717/peerj-cs.3504.

European Union. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), article 27: Fundamental rights impact assessment for high-risk ai systems, 2024.

Farheen, S., Cheema, A. A., Ullah, R. S., and Bandeali, M. M. Equity and bias in ai educational tools: A critical examination of algorithmic decision-making in classrooms. *The Critical Review of Social Sciences Studies*, 3 (3):67–85, 2025. doi: 10.59075/zqmnpa62.

Feng, S. et al. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Fricker, M. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.

Hulus, A. A systematic response to ethical blind spots in ai education: cross-cultural insights and the role of digital literacy. *Smart Learning Environments*, 13(1), 2026. doi: 10.1186/s40561-026-00437-1.

Judijanto, L. Beyond access: Cultural, ethical, and infrastructural challenges of ai in marginalised education contexts. *European Journal of Contemporary Education and E-Learning*, 3(6):83–98, 2025. doi: 10.59324/ejceel.2025.3(6).07.

Kasirzadeh, A. Value pluralism and ai value alignment. In *Advances in Neural Information Processing Systems*, 2024.

Kay, J., Kasirzadeh, A., and Mohamed, S. Epistemic injustice in generative ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2024.

Khan, S. The ethical imperative: Addressing bias and discrimination in ai-driven education. *Social Sciences Spectrum*, 2023.

Kidd, C. and Birhane, A. How ai can distort human beliefs. *Science*, 380(6651):1222–1223, 2023.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: 10.1073/pnas.1915768117.

Kohnke, L. and Foung, D. Deconstructing the normalization of data colonialism in educational technology. *Education Sciences*, 14(1):57, 2024. doi: 10.3390/educsci14010057.

Kukutai, T. and Taylor, J. *Indigenous data sovereignty: Toward an agenda*. ANU press, 2016.

Li, C., Zhang, H., Xu, Y., Xue, H., Ao, X., and He, Q. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. ACL 2025, 2025.

Mantelero, A. The fundamental rights impact assessment (fria) in the ai act: roots, legal obligations and key elements for a model template. *SSRN Electronic Journal*, 2024. doi: 10.2139/ssrn.4782126.

Mayer, A. Infrastructuring contestability: A framework for community-defined ai value pluralism. arXiv preprint, 2025.

Mayfield, E. et al. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 444–460, 2019. doi: 10.18653/v1/W19-4446.

Mishra, A. Ai alignment and social choice: Fundamental limitations. arXiv preprint, 2023.

Mollema, J. T. M. A taxonomy of epistemic injustice in ai and the case for generative hermeneutical erasure. *arXiv preprint arXiv:2504.07531*, 2025.

Mouta, A., Pinto-Llorente, A. M., and Torrecilla-Sánchez, E. M. Uncovering blind spots in education ethics: Insights from a systematic literature review on artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, pp. 1–40, 2024. doi: 10.1007/s40593-023-00384-9.

Mumtaz, S., Carmichael, J., Weiss, M., and Nimon-Peters, A. Ethical use of artificial intelligence based tools in higher education: are future business leaders ready? *Education and Information Technologies*, 30:7293–7319, 2025. doi: 10.1007/s10639-024-13099-8.

Nyaaba, M., Wright, A. L., and Choi, G. L. Generative ai and digital neocolonialism in global education: Towards an equitable framework. *ArXiv*, abs/2406.02966, 2024. doi: 10.48550/arxiv.2406.02966.

Pekrun, R. et al. Emotions associated with receiving grades: A control-value theory perspective. *Educational Psychology Review*, 2023.

Porayska-Pomsta, K., Holmes, W., and Nemorin, S. The ethics of ai in education. *ArXiv*, abs/2406.11842, 2023. doi: 10.48550/arxiv.2406.11842.

R, S. B., S, S. M., Min, J., D, D., and Sakkan, T. Decolonizing the digital classroom: A critical analysis of power, privilege, and algorithmic bias in ai-mediated learning environments. *Asian Journal of Interdisciplinary Research*, pp. 301–330, 2025. doi: 10.54392/ajir25417.

Russo Carroll, S. et al. The care principles for indigenous data governance. *Data Science Journal*, 19(1), 2021.

Samuel, Y. et al. Cultivation of human centered artificial intelligence: culturally adaptive thinking in education (cate) for ai. *Frontiers in Artificial Intelligence*, 6, 2023. doi: 10.3389/frai.2023.1198180.

Schiff, D. S. Education for ai, not ai for education: The role of education and ethics in national ai policy strategies. *International Journal of Artificial Intelligence in Education*, 32(3):527–563, 2022. doi: 10.1007/s40593-021-00270-2.

Sloane, M., Moss, E., Awomolo, O., and Forlano, L. Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–6, 2022.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070, 2024.

Srewa, M., Zhao, T., and Elmalaki, S. Pluralistic alignment in llms via federated learning. arXiv preprint, 2025.

Srewa, M., Zhao, T., and Elmalaki, S. Appa: Adaptive preference pluralistic alignment for fair federated learning. arXiv preprint, 2026.

Ter-Minassian, L. Democratizing ai governance: Balancing expertise and public participation. arXiv preprint, 2025.

Ullstein, C., Jarvers, S., Hohendanner, M., Papakyriakopoulos, O., and Grossklags, J. Participatory ai and the eu ai act, 2025.

Vamplew, P., Hayes, C. F., Foale, C., Dazeley, R., and Harland, H. Multi-objective reinforcement learning: A tool for pluralistic alignment. arXiv preprint, 2024.

Xiong, N. and Singh, A. Projection optimization: A general framework for multi-objective and multi-group rlhf. arXiv preprint, 2025.

Yang, W. and Al-Masri, E. Democratizing differential privacy: A participatory ai framework for public decision-making. arXiv preprint, 2025.

Yoder-Himes, D. R. et al. Racial, skin tone, and sex disparities in automated proctoring software. *Frontiers in Education*, 7, 2022. doi: 10.3389/feduc.2022.881449.

Zeide, E. and Nissenbaum, H. Learner privacy in moocs and virtual education. *Theory and Research in Education*, 16(3):280–307, 2018.

Zhou, Z., Liu, J., Shao, J., Yue, X., Yang, C., Ouyang, W., and Qiao, Y. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. arXiv preprint, 2024.

A. Localized FRIA Checklist for EdTech AI

Table 2. Localized FRIA Checklist for EdTech AI

Audit dimension	**What to test**	**Example local subgroup/context**	**Failure signal**	**Mitigation**
Subgroup performance	Standard metrics (accuracy, F1) disaggregated by intersectional groups	Non-standard dialect speakers, religious minorities	Sharp performance drop for specific group	Federated finetuning on local data
Infrastructure stress	Performance under low bandwidth, poor lighting, shared devices	Rural schools / low-resource households	Sharp error-rate rise, false misconduct flags	Disable feature, lower-risk fallback mode
Context sensitivity	Alignment of output with local cultural norms, pedagogical values	Community-specific history or ethical scenarios	Generation of offensive/irrelevant content	Steerable response policies with local review
Contestability / appeals	Efficiency, transparency, and outcomes of human review process	Students appealing automated grading/proctoring flags	High override rate by human review, opaque decisions	Formalized teacher override and feedback loops
Post-deployment monitoring	Continued drift of model and community impact over time	Evolving local dialects or curriculum shifts	Degrading fairness metrics, student self-censorship	Periodic re-audit, localized deployment restrictions