Vol. 10, Issue 4, Pp 15-34, Autumn 2026

Socio-Spatial Studies

Journal of Socio-Spatial Studies

Homepage: https://soc.gpmsh.ac.ir

Original Research Paper

Training Locally, Thinking Globally: The Political Economy of Community AI in Television and Satellite Broadcasting and the Structural Limits of Computational Decentralization

Haider Falah Zaeid[1]

[1] *Assistant Professor, College of Media, University of Thi Qar, Thi Qar, Iraq*

ARTICLE INFO

Keywords:
artificial intelligence
satellite broadcasting
NWICO
data colonialism
data sovereignty
compute divide
broadcast archives
low-resourced languages
political economy of technology

Received: Jun. 10, 2026
Revised: Jul. 19, 2026
Accepted: Aug. 27, 2026
Published: Oct. 1, 2026

ABSTRACT

Across the Global South, and within Indigenous and minority-language communities worldwide, a proposal has gained currency: if our languages, our knowledge, and our everyday expression are being harvested to train artificial intelligence systems owned and operated elsewhere, then we should train such systems ourselves. This article examines the political economy of that proposal, and argues that broadcasting is the field in which it is best understood. The argument has two parts. First, broadcasting supplies the historical precedent: the demand for a New World Information and Communication Order, and the satellite sovereignty controversies of the 1970s and 1980s, posed the same structure of claim, and the outcome was named by Straubhaar as asymmetrical interdependence — genuine local production capacity coexisting with persistent structural subordination. Second, broadcasting supplies the material substrate: for many low-resourced languages the largest existing body of recorded speech is held not on the open web but in the archives of radio and television stations, which makes broadcasters the effective gatekeepers of the corpora on which community AI depends. Drawing on the literature on data colonialism, the documented "compute divide," recent feasibility modelling of sovereign-scale training, and community-led initiatives including the Masakhane network and the Māori broadcaster Te Hiku Media, the article contends that the binding constraints are less about where computation physically happens than about who governs four resources: data, compute and infrastructure, frameworks and base models, and the distribution of resulting value. It specifies four conditions under which local training would constitute genuine rather than nominal sovereignty, argues that each is political and economic rather than technical, and identifies public service and community broadcasters as institutional actors whose archives, mandates and governance structures make them overlooked candidates for the role.

INTRODUCTION

A recurring proposition has emerged across the Global South and within Indigenous and minority-language communities worldwide: if our languages, our knowledge, and our everyday expression are being harvested to train artificial intelligence systems owned and operated elsewhere, then we should train such systems ourselves. We should build models in our own languages, on corpora we control, reflecting our values and serving our

* Corresponding Author
Haiderfalh91@utq.edu.iq
https://orcid.org/0000-0002-4005-6880

How to Cite this article:
Falah Zaeid, H. (2026). Training locally, thinking globally: The political economy of community AI in television and satellite broadcasting and the structural limits of computational decentralization. *Socio-Spatial Studies*, 10(4), 15-34.
https://doi.org/10.22034/soc.2026.256632
URL: https://soc.gpmsh.ac.ir/article_256632.html

Copyright © The Author(s);
This is an open access article distributed under the terms of the The Creative Commons Attribution (CC BY 4.0) license:
https://creativecommons.org/licenses/by/4.0/deed.en

priorities, and we should do so on our own territory and under our own governance. The proposition is intuitively compelling. It responds to a visible asymmetry in which the linguistic and cultural production of billions of people is converted into commercial products that are then sold back to those same populations, and it promises to reverse that flow by relocating both the activity and the value of model-building closer to the communities whose data make it possible.

This proposition has moved well beyond rhetoric. The Masakhane network has coordinated a continent-spanning, community-driven effort to build natural language processing resources for African languages, demonstrating that distributed, participatory research can produce datasets, benchmarks, and trained models for languages that commercial systems have long neglected (Nekoto et al., 2020). In South Asia, sustained initiatives have assembled monolingual corpora, evaluation benchmarks, and pre-trained multilingual models for Indian languages spoken by hundreds of millions of people but underserved by English-centric systems (Kakwani et al., 2020). Across Latin America, coordinated shared tasks have built translation resources for Indigenous languages including Aymara, Quechua, Guarani, and Shipibo-Konibo (Mager et al., 2021). In Aotearoa New Zealand, the Māori radio and television broadcaster Te Hiku Media has built speech recognition for te reo Māori from audio gathered within its own community and released under a licence of its own devising (Hao, 2022; Te Hiku Media, 2022). Governments, meanwhile, increasingly frame data localization, public compute investment, and "sovereign AI" as instruments of national technological policy.

This article takes that intuition seriously while subjecting it to critical scrutiny, and it argues that the field in which the proposition is best understood is broadcasting. That claim has two parts, corresponding to two roles broadcasting plays in the argument.

The first is historical. The demand now being made of artificial intelligence was made, in almost identical structural terms, of satellite television. Between the late 1960s and the early 1980s, the prospect of broadcasting directly across borders provoked an international controversy about whether the free flow of information was a liberty to be protected or an asymmetry to be regulated, and the demand that crystallized around it — the New World Information and Communication Order — proposed the same remedy now proposed for AI: build local capacity, assert authority over what enters, redistribute the means of production. That episode has an outcome, and the outcome is instructive. Local and regional production capacity did grow, substantially; what it did not produce was autonomy. Straubhaar (1991) named the result asymmetrical interdependence, and the term remains the most useful inheritance the present debate could take from the older one. Section 3 develops this precedent, because an argument about the limits of local production should begin by noticing that those limits have been mapped before.

The second role is material, and it concerns where the training data actually are. For many low-resourced languages, the largest existing body of recorded speech is not on the open web at all; it sits in the archives of radio and television stations that have been recording in those languages, daily, for decades. Recent work in speech technology has begun to exploit precisely this, assembling usable corpora from donated broadcast archives in languages for which no other substantial holdings exist (Doumbouya, Einstein, & Piech, 2021). The consequence for the political economy of community AI is direct: the broadcaster is the effective gatekeeper of the corpus, and the question of who governs training data for a minority language is, in many cases, a question about broadcast archives and the institutions that hold them. That the most fully realized example of community-governed language technology anywhere — Te Hiku Media — is itself a broadcaster is not a coincidence.

The central claim of the article is that local training is necessary but not sufficient: it addresses real and important harms, but it confronts structural barriers that technical capacity and local effort, on their own, cannot overcome. The extractive logic of

contemporary AI development is not reducible to the geography of computation — to the question of where, physically, a model is trained. It is embedded in a wider circuit of ownership and control that determines who governs the data that become training corpora, who manufactures and allocates the specialized hardware on which training depends, who designs and maintains the software frameworks and releases the base models that most projects build upon, and who ultimately captures the economic and epistemic value that AI development generates. A community that trains a model on local data, but does so using imported hardware subject to foreign export controls, using frameworks developed and documented elsewhere, and by adapting a base model released under terms a distant corporation can change at will, has not exited the extractive circuit. It has relocated one stage of that circuit while remaining dependent on the others.

The argument therefore turns on a distinction between two senses of "local." In the first, thin sense, training is local if the computation occurs within a national border or under a local institution's nominal control. In the second, thick sense, training is local — and approaches genuine sovereignty — only when the community exercises meaningful authority over the full stack of resources on which the model depends, from corpus to compute to framework to the allocation of resulting benefit. Much of the policy enthusiasm for "sovereign AI" elides this distinction, treating a fundamentally political-economic problem as though it were primarily a technical one. The risk is not merely conceptual. To the extent that local training is presented as a self-sufficient solution, it can divert attention and resources from the structural questions on which the value of local training ultimately depends. That distinction is carried through the article as its organising device: each constraint section asks not whether a resource is available locally but who governs it.

The article proceeds as follows. After describing its method, Section 3 reconstructs the satellite broadcasting precedent. Section 4 characterizes the extractive circuit of foundation-model development through the lens of the data colonialism literature, and Section 5 examines what community-led initiatives have achieved. Sections 6 to 8 analyse the three principal material constraints in turn — data, compute, and frameworks-and-base-models. Section 9 names the structural paradox that results and consolidates the analysis in a table; Section 10 tests the argument against five objections, including the objection that the broadcasting analogy misleads; and Section 11 specifies the four conditions under which local training could constitute genuine rather than nominal sovereignty, and identifies broadcasters as candidate institutional stewards. The article concludes that the path forward lies not in choosing between local training and structural reform, but in pursuing both.

2. METHOD

This article employs critical literature analysis, an interpretive method oriented toward identifying structural patterns, conceptual tensions, and points of contradiction across bodies of scholarship that do not ordinarily speak to one another. Rather than measuring the outcomes of particular initiatives or conducting a systematic review with quantitative synthesis, the analysis brings into conversation literatures from several fields — the political economy of communication and broadcasting, data justice and data sovereignty studies, decolonial and postcolonial studies of computing, natural language processing for low-resourced languages, and research on the infrastructure and economics of machine learning. Bringing the first of these into contact with the last is the article's principal methodological move: the international communication scholarship on satellite broadcasting and the technical literature on low-resource language technology address structurally similar problems and almost never cite one another.

Sources were identified through structured searches of scholarly databases and indexes, including Google Scholar and disciplinary repositories such as the ACL Anthology and arXiv, supplemented by citation tracing from foundational works to recent contributions. Search terms combined conceptual vocabulary — "data colonialism," "algorithmic coloniality," "data sovereignty," "technological sovereignty," "compute divide," "NWICO," "media imperialism" — with domain-specific terms such as "low-resourced languages," "broadcast archives," "multilingual

language models," and "AI in the Global South." Priority was given to peer-reviewed articles, conference proceedings in computational linguistics, and scholarly monographs, alongside a small number of recent preprints where they reported directly relevant empirical modelling not yet available in the published literature. Two community initiatives — the Masakhane network and Te Hiku Media — are used throughout as running illustrations rather than as case studies in the methodological sense: they are drawn upon to make abstract constraints concrete, not analysed as data.

Claims of empirical fact are attributed to specific identifiable sources. Where a claim rests on industry documentation or reporting rather than on peer-reviewed research — as is unavoidable for the current structure of the accelerator market and for the practices of particular organizations — the source is named as such, and the claim is stated at the level of confidence the source supports. Where the available evidence is partial, contested, or rapidly changing, as it is especially for the economics of model training, the analysis flags that uncertainty rather than resolving it artificially.

The method has evident limitations. A literature analysis cannot establish causal claims, nor can it substitute for fieldwork with the communities whose initiatives it discusses; its conclusions are interpretive and provisional, intended to clarify structure rather than to forecast outcomes. The historical comparison in Section 3 is offered as an analytical precedent rather than as historiography, and readers with specialist knowledge of the NWICO period will find it compressed. The analysis is also necessarily selective, foregrounding the political economy of training over adjacent questions — model evaluation, downstream deployment, content moderation — that merit separate treatment. A further limitation deserves explicit statement: the article discusses community initiatives from outside them, on the basis of their published accounts, and its reading of what those initiatives demonstrate is therefore an outsider's reading, which practitioners might reasonably contest.

3. THE PRECEDENT: SATELLITE BROADCASTING AND THE UNFINISHED BUSINESS OF NWICO

The proposition that a community should build its own media technology rather than receive someone else's is not new, and the field that has studied it longest is broadcasting. Between the late 1960s and the early 1980s, the arrival of communication satellites — and in particular the prospect of direct broadcasting from satellite to home receiver — provoked an international controversy whose terms are startlingly close to the present one. Broadcast signals do not respect borders. A satellite positioned over one state can deliver programming into another without the second state's consent and without passing through any national gatekeeper. The question this raised, whether the free flow of information across borders was a liberty to be protected or an instrument of asymmetry to be regulated, dominated debate at UNESCO for more than a decade.

The demand that crystallized around this question was the New World Information and Communication Order. Its analysis, given its fullest statement in the MacBride Commission's report (International Commission for the Study of Communication Problems, 1980), was that the flow of news and programming ran overwhelmingly in one direction, from a small number of producing centres in the North to the rest of the world; that the infrastructure carrying that flow was owned and operated by the producing countries; and that the states on the receiving end had neither the capacity to produce at comparable scale nor any authority over the terms on which they received. The remedy proposed was, in its essentials, the remedy now proposed for artificial intelligence: build local capacity, assert authority over what enters, and redistribute the means of production so that the flow becomes two-way.

What followed is instructive precisely because it was neither a vindication nor a defeat. Local and regional production capacity did grow, substantially and in some cases spectacularly. Regional production centres emerged in Latin America, in Egypt and later across the Arab world, in India and

in East Asia; audiences proved to prefer culturally proximate programming when it was available to them; and by the 1990s the flow of television was visibly less unidirectional than the media imperialism thesis had assumed. Straubhaar (1991) captured this development with a term that is the most useful inheritance the present debate could take from the older one: asymmetrical interdependence. His argument was that the simple imperialism model was wrong, because production had genuinely dispersed and audiences genuinely exercised choice; but that what replaced it was not autonomy. It was a structured interdependence in which the weaker parties had real capacity and real preferences and remained, nonetheless, structurally subordinate — dependent on capital, formats, technology, and distribution arrangements they did not control.

The satellite era makes the point concretely. The Arab satellite broadcasting boom from the early 1990s produced channels of genuine regional reach and political consequence, established on the initiative of regional actors rather than Northern firms. But the transponders were leased or purchased; the uplink equipment, the encryption systems, and the receiving hardware were manufactured elsewhere; the advertising markets and the audience-measurement systems on which commercial viability depended were organized around standards set elsewhere; and the regulatory questions of orbital slots and spectrum were settled in international fora in which the producing states held the stronger hand (Sakr, 2001). Sovereignty over content was won. Sovereignty over the stack beneath the content was not, and was never seriously on offer.

Table 1 sets the two episodes side by side. The rows are the layers at which control is exercised; reading across them shows how closely the present configuration reproduces the earlier one, and where it does not.

Table 1. Two episodes of communicative sovereignty compared

	Satellite broadcasting (1970s–1990s)	**Foundation models (2020s)**
The demand	National and regional control over what is broadcast and what is received	Community and national control over models trained on local language and knowledge
Raw material	Programming, news, and cultural production	Corpora of text and recorded speech — much of the speech held in broadcast archives
Means of production	Studios and production capacity, which did genuinely localize	Training compute, partially attainable but on foreign-supplied hardware (Malagon et al., 2025)
Enabling infrastructure	Transponders, uplink equipment, encryption and receivers, foreign-manufactured; orbital slots allocated in international fora	Accelerators, frameworks and base models, foreign-designed and revocable at the originator's discretion
Commercial substrate	Advertising markets and audience measurement organized to externally set standards	Licensing terms, interface access and evaluation benchmarks set externally
Outcome	Asymmetrical interdependence (Straubhaar, 1991): real capacity, persistent structural subordination	The open question this article poses

Two conclusions follow, and they pull in different directions. The first is cautionary. The structure this article describes — local production coexisting with dependence at every constitutive layer — is not a novel feature of artificial intelligence. It is what happened the last time a communication technology

provoked a sovereignty demand, and the vocabulary for describing it has existed since 1991. That the AI debate has largely proceeded without reference to this precedent is itself a symptom of how readily each new technology is treated as unprecedented.

The second is more encouraging. If the satellite era shows that building local capacity does not by itself deliver sovereignty, it also shows that building local capacity is not futile. Arab satellite broadcasting changed what could be said and seen in the region; regional production in Latin America and South Asia reshaped what audiences watched and how they understood themselves. Those were real gains, achieved within a structure that remained asymmetrical. The lesson is not that the effort was pointless but that its achievements and its limits must be named separately — which is the discipline this article attempts to apply to the case of AI.

Recent work has begun to recover the NWICO framework explicitly for the digital era, arguing that neoliberal globalization did not dissolve the imbalances the MacBride Commission identified but reconstituted them around platform expansion and data extraction, with ownership of data, algorithms, governance systems, and infrastructure remaining concentrated (Joynes, Dwivedy, & Kumar, 2026). The present article can be read as a contribution to that recovery, applied to the specific case of model training, and conducted with the benefit of what the satellite episode taught about the distance between capacity and control.

4. THE EXTRACTIVE CIRCUIT: DATA COLONIALISM AND THE FOUNDATION-MODEL ECONOMY

If the satellite episode supplies the precedent, the data colonialism literature supplies the vocabulary for the present case. Couldry and Mejias (2019) advance the thesis that contemporary data practices constitute a new colonial appropriation: just as historical colonialism annexed land, bodies, and natural resources as raw material for capital, the data economy annexes human life itself — expressed as digitized behaviour, communication, and cultural production — as the raw material of a new round of accumulation. The argument is not that the data economy resembles colonialism metaphorically, but that it instantiates a structurally analogous appropriation, establishing what they term "data relations" through which human experience is rendered extractable and converted into value for those who control the means of processing.

Foundation-model development extends and intensifies this dynamic in a specific way. Where earlier data extraction largely harvested the ongoing behavioural traces of users — clicks, queries, locations, dwell times — large language model training reaches backward and outward into the accumulated archive of recorded human expression. Digitized books, the open web across millions of domains, code repositories, transcribed speech, encyclopedias, and the written output of countless communities become the corpus from which models are built. The decolonial analysis of AI developed by Mohamed, Png, and Isaac (2020) is particularly useful here, because it specifies the mechanisms through which this produces harm. They identify a triad of what they call algorithmic coloniality: algorithmic oppression, in which systems entrench existing hierarchies; algorithmic exploitation, in which the labour and data of the dispossessed are appropriated under conditions they do not control; and algorithmic dispossession, in which value and decision-making authority are concentrated in a few centres at the expense of the many. Crucially, Mohamed and colleagues frame these not as incidental defects but as expressions of a power structure with deep historical roots.

Birhane (2020) sharpens the political reading. Examining the importation of Western-designed digital systems into African contexts, she argues that algorithmic colonialism differs from its historical antecedent in agency more than in form: where traditional colonialism was driven principally by states and their militaries, algorithmic colonialism is driven principally by corporations and their commercial agendas, operating not through overt force but through control of digital infrastructure, platforms, and standards. This is a significant departure from the satellite era, in which the contending parties were states and the arena was intergovernmental. The consequence is that

technologies developed with Northern assumptions, priorities, and commercial models are deployed in the South with little regulation or critical scrutiny, producing benefits that accrue disproportionately to the originating firms. Birhane's response is not technophobic; she argues instead for AI rooted in local community needs and interests, and points explicitly to community-driven efforts as a counter-model.

The extractive circuit, then, has a recognizable shape. Expression and data are gathered, at vast scale and frequently without meaningful consent, from populations distributed across the globe. The conversion of that raw material into valuable systems — the training itself — is concentrated at a small number of sites that command the necessary capital, expertise, and computational infrastructure. The resulting products are then distributed back to the originating populations as commercial services, typically by subscription, so that communities whose linguistic and cultural production helped constitute the training corpus must purchase access to systems built upon it. Raw material flows inward to centres of processing; finished goods and the profits from them flow back outward and upward. The reader who has followed Section 3 will recognize the shape: it is the one-way flow the MacBride Commission described, reconstituted around a different commodity. What is novel in the AI case is the nature of the raw material — the accumulated intellectual, linguistic, and cultural inheritance of humanity — and the nature of the finished good — systems that increasingly mediate access to information, employment, education, and public life. Research on virtual education also illustrates how digital platforms can expand access to educational resources for geographically and socially underserved groups. Dastyar et al. (2023) highlight the potential of virtual learning to facilitate collaboration, knowledge exchange, and cross-cultural understanding, demonstrating the importance of accessibility and inclusion in digital learning environments.

The growing scholarly consensus that these dynamics constitute a decolonial problem for data and technology research (Couldry & Mejias, 2023) reflects the recognition that the stakes are not merely economic but epistemic, bearing on whose knowledge is represented and whose is rendered marginal.

5. THE PROMISE OF LOCAL TRAINING: COMMUNITY INITIATIVES AND WHAT THEY ACHIEVE

Before turning to constraints, it is essential to take the affirmative case seriously, because the achievements of community-led AI initiatives are real, hard-won, and too easily minimized by a purely structural critique. The most instructive example is the Masakhane network, whose participatory approach to machine translation for African languages has been documented in detail by Nekoto et al. (2020). Their account describes a research effort that is deliberately distributed, open, and online, mobilizing more than four hundred participants across more than twenty countries to build translation datasets and benchmarks for over thirty African languages. The methodological argument is as important as the artifacts produced. Nekoto and colleagues contend that "low-resourcedness" is not merely a technical condition — a shortage of digitized text — but a reflection of systemic inequities in whose languages have been documented, funded, and deemed worthy of investment. Their response is participatory research that enables people without formal training in machine learning to make genuine scientific contributions, thereby redistributing not only the outputs of research but the capacity to produce it.

A second initiative illustrates a different dimension of the same project and recurs throughout this article. Te Hiku Media is a Māori radio and television broadcaster in the far north of Aotearoa New Zealand, established in 1990 and committed to the revitalization of te reo Māori. Its move into language technology grew directly out of that broadcasting function: a station that has been recording, transmitting, and archiving speech in a minority language for three decades is, by that fact, the custodian of one of the largest holdings of that language in existence. Rather than scraping available audio, it began by seeking the support of elders, then ran a community crowdsourcing campaign that gathered several hundred hours of labelled speech from thousands of contributors in a matter of days.

Crucially, it released the resulting resources not under a conventional open licence but under a licence of its own devising, the Kaitiakitanga License, which grants access only to those who agree to respect Māori values, remain within the bounds of the consent given, and pass benefits derived from the data back to Māori (Hao, 2022; Te Hiku Media, 2022). Te Hiku's work is instructive precisely because it is among the most thoroughly realized examples of community governance anywhere in language technology; what happens to its dependencies, examined in Sections 7 and 8, is therefore a demanding test of the argument advanced here.

This redistribution of capacity is a benefit distinct from, and arguably more durable than, any particular model. When a continent-wide network trains researchers, establishes collaborations across institutions and borders, builds shared datasets, and creates pathways for local scholars to participate in and shape AI research, it constructs institutional infrastructure that did not previously exist. Birhane (2020) singles out exactly this kind of community-driven work as a counter to algorithmic colonialism, precisely because it locates agenda-setting within the affected communities rather than outside them. The value here is not only instrumental but constitutive: it changes who counts as a producer of technological knowledge. Dariush et al. (2025) propose a participatory educational framework that brings together traditional ecological wisdom, contemporary learning approaches, and digital technologies to support collective awareness and community engagement. The parallel with the satellite era is close, and worth stating: the enduring achievement of regional broadcasting was less any particular programme than the production capacity, the trained workforce, and the institutions that outlasted it.

The initiatives also address a concrete and consequential failure of dominant systems: their poor performance in most of the world's languages. The political economy of NLP has produced a stark stratification, analysed by Joshi et al. (2020), in which a handful of resource-rich languages — English foremost among them — command the overwhelming share of data, tooling, and research attention, while the majority of the world's languages remain "have-nots," poorly served or entirely absent from language technology. The consequences are not abstract. Translation systems for low-resource language pairs frequently produce unreliable output; speech recognition underperforms for the accents and languages of the underserved; content moderation systems trained predominantly on English mishandle other languages, both failing to detect harm and suppressing legitimate expression. Noble (2018) has shown, in the adjacent domain of search, how systems trained on skewed data reproduce and amplify representational harm; the same logic applies to language models whose training data underrepresent most of humanity. Models trained deliberately on local-language data can materially outperform generic multilingual systems for the communities they serve, and the downstream benefits are substantial — for access to education and public information in one's strongest language, for the documentation and revitalization of endangered languages, and for financial inclusion, where the capacity of institutions to serve customers in their own languages is one of several factors bearing on access to services (Habib Zadeh Khiyaban & Sabbar, 2026).

There is, moreover, a normative dimension that community governance brings into focus. Indigenous data sovereignty frameworks insist that data about and from a community belong, in a meaningful sense, to that community, and that decisions about their use are properly exercised by the community rather than appropriated by external actors (Kukutai & Taylor, 2016). These commitments have been articulated as a set of principles for Indigenous data governance — collective benefit, authority to control, responsibility, and ethics — developed through international consultation and framed explicitly as a complement to, and a corrective for, the purely technical emphasis of open-data norms (Carroll et al., 2020). Recent work has translated such principles into concrete design and evaluation criteria; Munn (2024), for instance, articulates a set of tests for AI systems derived from Indigenous Māori principles, illustrating how community values

can be operationalized rather than merely invoked. Local training, undertaken under such governance, is therefore not only a technical strategy for better performance but an assertion of authority over linguistic and cultural resources. These achievements are genuine and deserve sustained support. The argument of this article is not that they are illusory, but that, on their own, they encounter limits set by parts of the system that lie beyond local control.

6. THE FIRST CONSTRAINT: BROADCAST ARCHIVES, DATA AVAILABILITY, AND GOVERNANCE

The first constraint on local training is the most easily underestimated, because it appears at first to be a problem of effort rather than of structure. To train a language model, a community needs a corpus: a large, diverse, reasonably clean body of text or speech in the target language. The obvious source is the open web, together with digitized books and other public repositories. But the composition of what is available online is not neutral; it is itself a sediment of historical and ongoing inequality. The web overrepresents a small number of languages and the regions, perspectives, and knowledge systems associated with them, while most of the world's languages are sparsely present or effectively absent. For many languages, the quantity of digitized text simply does not exist at the scale that contemporary methods assume. The very condition that local training seeks to remedy — the marginalization of a language in the digital sphere — is also a precondition that makes local training difficult.

That statement of the problem is standard, and it is incomplete, because it assumes the web is where one should be looking. For spoken language it frequently is not. Radio and television stations in linguistically diverse regions have, in many cases, been recording daily in languages the web barely represents — news, interviews, call-in programmes, drama, religious and agricultural broadcasting — and have accumulated audio over decades at a scale no research project could replicate. Doumbouya, Einstein, and Piech (2021) demonstrated what follows from this. Working with six Guinean radio stations that donated their archives, they assembled the West African Radio Corpus, some 142 hours of audio across more than ten languages, and showed that a speech encoder trained on this noisy broadcast material performed comparably on West African speech recognition tasks to one trained on a high-quality English corpus nearly seven times larger — producing, in the process, the first usable speech recognition models for Maninka, Pular, and Susu. Comparable corpora have since been assembled from broadcast archives elsewhere in Africa, and reviews of the field now treat broadcast and media archives as a distinct source category alongside academic collections and research programmes.

This reframes the data constraint rather than dissolving it. If broadcast archives are, for many languages, the largest available body of recorded speech, then the broadcaster is the gatekeeper of the corpus, and the question of who governs training data becomes a question about broadcasting institutions and the terms on which they hold their holdings. That question has no general answer. A state broadcaster in an authoritarian setting, a commercial satellite channel with rights obligations to producers and performers, a community radio station, and a public service broadcaster with a statutory remit are four quite different custodians, with different incentives and different accountabilities to the language communities whose speech they hold. It is worth registering the verb in the Doumbouya study: the stations donated their archives. Nothing compelled them to, and nothing would have prevented them from selling the same material to a well-resourced foreign firm instead. The corpus existed because institutional gatekeepers chose to release it under terms favourable to a public-interest project, which is a governance outcome and not a technical one.

Communities that cannot draw on such archives must create data, and here the constraint shifts from availability to labour and governance. Building a corpus through fieldwork, digitization of historical materials, transcription, and annotation is extraordinarily labour-intensive and demands sustained institutional and financial support of a kind that is itself unevenly distributed. The participatory model documented by Nekoto et al. (2020) is in part a response to exactly this problem. Te Hiku Media's crowdsourcing campaign is

another, and its speed — thousands of contributors within days — is a measure of what community legitimacy can mobilize that a commercial data-collection contract cannot. Yet reliance on such effort raises its own questions of sustainability and fairness, given what the broader literature has documented about the conditions of data work. Gray and Suri (2019) describe the vast and largely invisible human labour, much of it performed in the Global South under precarious conditions and minimal pay, that underwrites systems marketed as automated. The danger is that data production for "local" models, if not deliberately structured around fair compensation and genuine community control, can reproduce the very extractive patterns it purports to escape. That this tension is publicly felt, and not merely an academic observation, is visible in the discourse that has formed around artificial intelligence and the displacement of work, where the question of who bears the costs of automation and who captures its gains is argued out in explicitly moral terms (Sabbar & Habib Zadeh Khiyaban, 2023).

This is where data sovereignty becomes more than a slogan. The questions that determine whether corpus-building is empowering or extractive are governance questions: who owns the resulting dataset; who decides how it may be used, and for how long; whether the community retains authority over derivative uses and downstream models; and whether participation yields benefit to the community or merely renders its language legible for appropriation by better-resourced actors. Indigenous data sovereignty frameworks (Kukutai & Taylor, 2016), the principles of collective benefit and authority to control articulated in the CARE framework (Carroll et al., 2020), and their operationalization in concrete principles for AI (Munn, 2024) provide normative resources for answering these questions. The Kaitiakitanga License represents an unusually developed attempt to give such principles legal form, binding access to a commitment that benefits return to the community whose data are used — and it is, again, a broadcaster that devised it. But the instrument also reveals the gap between normative architecture and enforceable practice: its adoption beyond its originators has been limited, and questions about its enforceability against a determined or distant infringer remain open (Hao, 2022). Under prevailing norms and law, meanwhile, material that appears online is treated as available for the taking unless specifically protected; the burden falls on communities to defend their data rather than on collectors to justify extraction. Until that default is reversed, corpus-building for local models proceeds on contested ground, and the mere fact that training happens locally does nothing to settle the underlying question of who governs the data.

7. THE SECOND CONSTRAINT: COMPUTE AND INFRASTRUCTURE

The second constraint is the most material and, in the public imagination, the most decisive: the computational infrastructure on which model training depends. Training a large model requires substantial quantities of specialized processors operating continuously for extended periods, supported by reliable electrical power, cooling, networking, and the engineering expertise to run them. These resources are distributed with extreme inequality, and that inequality has been documented as a structural feature of the field rather than a transient condition. Ahmed and Wahed (2020), analysing a dataset of 171,394 papers from fifty-seven major computer science venues, demonstrate the emergence of what they term a "compute divide." Following the rise of deep learning around 2012, participation in elite AI research became increasingly concentrated among large technology firms and a small set of highly ranked universities, crowding out mid-tier and lower-tier institutions. The mechanism they identify is precisely access to computational resources. The same authors, in subsequent work, document the growing influence of industry across AI research more broadly, as firms come to command disproportionate shares of compute, data, and talent (Ahmed, Wahed, & Thompson, 2023).

The compute divide has a pronounced geographic dimension, and it is the closest analogue in the present case to the transponder problem of the satellite era. The manufacture of advanced AI accelerators is dominated by a very small number of

firms, and the supply of these chips is shaped by export-control regimes set by the governments of producing countries, principally the United States. The degree of concentration is evident in the structure of the planning literature itself: feasibility studies of sovereign training are conducted by comparing successive generations of a single vendor's accelerators and by checking the resulting configurations against export-control thresholds set elsewhere (Malagon et al., 2025). Communities and states seeking to build training capacity may therefore find their access constrained not by price alone but by policy decisions made in pursuit of geopolitical objectives unrelated to their needs. Crawford (2021) situates this within a broader account of AI as a fundamentally extractive industry, one whose apparent immateriality conceals heavy dependence on minerals, energy, water, logistics, and labour — a planetary infrastructure whose costs and control are distributed as unequally as its benefits. On this view, the "cloud" is not a dematerialized abstraction but a concrete arrangement of mines, fabrication plants, data centres, and power grids, the commanding heights of which remain concentrated in a few hands.

The dependence is not confined to states operating at national scale. Te Hiku Media's speech models, built on audio the organization gathered and governs under its own licence, were nonetheless trained on accelerators designed and manufactured by the dominant foreign vendor. The most complete example of community data sovereignty in language technology thus runs, at the level of silicon, on hardware over which the community has no authority whatever. This is not a criticism of Te Hiku, which could hardly have done otherwise; it is an illustration of how far the governance of a corpus can be secured while the governance of computation remains entirely out of reach — the same distribution of wins and losses that the satellite broadcasters found, one technological generation earlier.

Yet the most recent evidence complicates the simplest version of the impossibility narrative, and intellectual honesty requires engaging it. A 2025 feasibility study by Malagon et al. modelled the technical and fiscal requirements of sovereign-scale language model training in Brazil and Mexico, varying accelerator generation and training duration across a dual-axis design. Its findings are striking: training a ten-trillion-token model was estimated to cost on the order of eight to fourteen million US dollars using current-generation accelerators, and nineteen to thirty-two million using a prior generation, with all of the modelled configurations remaining below both export-control thresholds and the limits of existing electrical infrastructure. The decisive variable, the authors argue, was not raw budget or even export policy but hardware efficiency; and they propose that extending training timelines can serve as a deliberate policy lever, allowing middle-income countries to produce usable, auditable, and locally aligned models by trading time for cutting-edge capacity, provided they do not insist on competing at the global frontier (Malagon et al., 2025).

This evidence cuts in two directions. On one hand, it punctures a fatalism that treats frontier-scale compute as simply unattainable outside a handful of countries: for non-frontier models adequate to many real linguistic and civic purposes, sovereign-scale training appears more achievable than commonly assumed. It is worth noting that the study's finding that export controls are not the binding constraint cuts against a common assumption in the critical literature, including assumptions this article might otherwise have made. On the other hand, the very framing of the finding underscores the depth of dependence. Feasibility is contingent on purchasing accelerators from a foreign vendor whose supply is governed by foreign export policy; on the efficiency of hardware designed elsewhere; and on a willingness to forgo the frontier, accepting models that are strategically sufficient rather than competitive with the systems produced by the dominant firms. To be below a threshold is still to be measured against a threshold one did not set. The costs cited, while not astronomical, remain substantial relative to the budgets of many states and represent real opportunity costs against other public priorities. And the analysis pertains to middle-income countries with relatively developed infrastructure; for lower-income settings, and for communities rather than states, the barriers are

correspondingly higher. The lesson is not that compute is an absolute barrier — it is not — but that even where training is feasible, it remains structurally dependent on resources whose production and governance lie outside the community's control. Feasibility is not the same as sovereignty.

8. THE THIRD CONSTRAINT: FRAMEWORKS, BASE MODELS, AND DEPENDENT DEVELOPMENT

The third constraint is less visible than hardware but no less consequential: the software frameworks and pre-trained base models that mediate nearly all contemporary model development. Training is conducted using a small number of dominant frameworks that implement the underlying algorithms, manage the distribution of computation across many processors, and coordinate the interaction between software and hardware. These frameworks are, in important respects, genuinely open, and their openness is a real public good. But the ecosystem around them — the documentation, the tutorials, the accumulated practical knowledge, the communities of practice that diagnose and fix problems — has been built largely in the Global North and overwhelmingly in English. A developer working in an underserved language community thus builds atop foundations designed for other contexts and conducts their work in a technical lingua franca not their own. Te Hiku Media again illustrates the point: its speech models were built using a toolkit released by the same vendor that manufactures the accelerators they run on. Openness at the level of source code coexists, here, with dependence at the level of who decides what the toolkit does and whether it continues to be maintained. The structural position is that of a broadcaster who owns the studio and the programming but not the transmission standard.

More consequential still is the widespread reliance on base models released by large firms. Because training a capable model from scratch is expensive in exactly the ways the previous section described, most community and national initiatives proceed instead by adapting an existing pre-trained model — fine-tuning it on local data to specialize it for a language or task. This is an eminently sensible strategy that places useful capability within reach of actors who could not train a frontier model independently, and it accounts for much of the practical progress in language technology for underserved languages. Collaboratively governed multilingual models developed through open scientific consortia, of which the BLOOM model produced by the BigScience workshop is the most prominent, demonstrate that genuine alternatives to purely proprietary provision are possible (BigScience Workshop et al., 2023), and Birhane (2020) identifies community-driven development as a counter-model to proprietary control.

But adaptation of a base model is also a relationship of dependence whose terms the adapting community does not set, and two recent lines of work give this claim an evidential basis it has often lacked. The first concerns what "open" actually means in this domain. Liesenfeld and Dingemanse (2024) assess generative AI systems against a framework distinguishing fourteen dimensions of openness — from training data and documentation to licensing and access method — and find that openness is a gradient rather than a binary, and that some of the most prominently marketed "open" releases score poorly across it. The practical implication for a community deciding what to build upon is considerable: a licence that permits fine-tuning tells one very little about whether the training data can be inspected, whether the documentation supports independent evaluation, or whether the terms will hold. The second concerns what openness accomplishes even where it is real. Widder, West, and Whittaker (2023) argue that openness alone does not disperse concentrated power in AI, because the resources that actually determine who can build — data at scale, compute, and the labour of a specialized workforce — remain concentrated regardless of how a model is licensed. Openness, on this account, can even consolidate incumbency, by making a firm's architecture the substrate on which an ecosystem of dependents is built.

The consequences for the adapting community follow directly. The capabilities, linguistic coverage, and embedded values of the base model are

determined by its originator; the license under which it is released, and the conditions of its continued availability, can be altered unilaterally; and a community that has invested heavily in building atop a particular model is exposed to decisions about support, pricing, terms of use, or discontinuation made by a distant firm whose commercial interests need not align with the community's. The appearance of autonomy that fine-tuning confers can therefore mask a deepening structural dependence, in which the most basic parameters of what a "local" model can do are inherited from a system the community neither built nor governs. Taken together, the three material constraints describe a development process that may be local at the point of training while remaining dependent at every other point.

9. THE STRUCTURAL PARADOX: DECENTRALIZED IN FORM, CENTRALIZED IN SUBSTANCE

The preceding analysis yields a paradox at the heart of the local-training project. The extractive circuit is not constituted solely, or even primarily, by the geographic location of computation. It is constituted by relations of ownership and control that run through the entire process: control over the data that become corpora, over the hardware and infrastructure on which training runs, over the frameworks and base models that mediate development, and over the distribution of the value that results. Relocating the single stage of training to local soil, while leaving the surrounding relations intact, does not exit the circuit. It produces a development process that is decentralized in form but remains centralized in substance. This is, in the vocabulary Section 3 recovered, asymmetrical interdependence: real capacity, genuinely held, within a structure one does not govern.

Table 2 consolidates the analysis by setting out each of the four layers alongside what is at stake in it, the constraint that binds it, and the condition that would have to be satisfied for authority over it to be genuine. Reading down the third column gives the diagnosis; reading down the fourth gives the programme, which Section 11 develops.

Table 2. Four layers of the AI development circuit: constraints and conditions for sovereignty

Layer	What is at stake	The binding constraint	Condition for genuine sovereignty
Data	Who governs the corpora from which a model is built, and on what terms	Digitised text is scarce for most languages; the largest speech holdings sit in broadcast archives under institutional rather than community control; prevailing norms presume extractability; corpus creation depends on labour that is easily under-compensated	Reversal of the default, so that community data may be used only under terms knowingly agreed and from which benefit flows back (Carroll et al., 2020; Kukutai & Taylor, 2016)
Compute and infrastructure	Who manufactures, allocates, and licenses the accelerators and the power, cooling and networking around them	Documented concentration of compute among firms and elite institutions (Ahmed & Wahed, 2020); accelerator manufacture dominated by one vendor and shaped by foreign export policy; feasible for middle-income states but not on their own terms (Malagon et al., 2025)	Computational infrastructure built and governed as a public good, with allocation outside the control of dominant firms
Frameworks and base models	Who designs the tooling, and who sets and may revoke the terms on which base models are available	Openness is a gradient rather than a binary, and the most visible releases score poorly on it (Liesenfeld & Dingemanse, 2024); openness alone does not disperse power (Widder et al., 2023); adaptation inherits coverage and values the adapter did not choose	Durable commons-based provision, governed through international and participatory processes and on terms no single firm may revoke
Value	Where the economic and epistemic returns of AI development	Returns accrue to model-builders, investors and a small technical workforce; the labour of data work is dispersed and poorly paid (Gray & Suri,	Mechanisms of reciprocity — compensation, shared ownership, technology-sharing obligations — that return value to contributing

Layer	What is at stake	The binding constraint	Condition for genuine sovereignty
	come to rest	2019); benefit-sharing instruments exist but are rare and of untested enforceability	communities

This is not an argument that local training is futile; the achievements documented earlier are real, and the feasibility evidence shows that meaningful local capacity is attainable. It is an argument about the level at which the binding problems sit. The danger is a species of technological solutionism: the framing of "we must train AI locally" can quietly recast a political-economic problem as a technical project of capacity-building, suggesting that with enough expertise, data, and effort the asymmetry will be overcome. Expertise, data, and effort are necessary and valuable. But no quantity of local technical capacity alters the facts that the manufacture of advanced accelerators is concentrated among very few firms and governed by foreign export policy; that the most capable base models are released, or withdrawn, at the discretion of a few organizations under terms they control; that the frameworks and the knowledge around them are shaped by communities elsewhere; or that, under prevailing norms, data are presumed extractable unless actively defended. These are matters of power, ownership, and governance, not of skill — which is precisely what the generation that built satellite channels and found itself still leasing transponders discovered.

The distinction drawn in Section 1 can now be stated precisely. Local training in the thin sense satisfies the first column of Table 2 and none of the fourth: computation occurs within a border, and the layers that determine what that computation can do remain governed elsewhere. Local training in the thick sense would satisfy the fourth column across all four rows. Every existing initiative sits somewhere between the two, and the interest of a case such as Te Hiku Media lies in how far it has travelled on the data row while remaining, of necessity, at the starting point on the others. Recognizing the paradox is the precondition for addressing it, because it directs attention from the question "how do we train locally?" to the prior and more difficult question "what would have to change for local training to constitute genuine sovereignty?"

10. OBJECTIONS AND REJOINDERS

Because the argument advanced here is, in part, a cautionary one, it is worth confronting directly the strongest objections to it, both to test its claims and to guard against misreading.

A first objection holds that the analysis overstates dependence. The widespread availability of open frameworks and open base models, the objection runs, already places real autonomy within reach. There is genuine force to this, and the article concedes it: openness is a real public good, and fine-tuning has placed useful capability within reach of actors who could not otherwise have obtained it. But the objection conflates availability with governance, and the evidence reviewed in Section 8 indicates that the conflation is consequential rather than merely conceptual. That a resource is described as open does not establish that its training data can be inspected or that its terms will hold (Liesenfeld & Dingemanse, 2024), and even where openness is genuine it does not by itself redistribute the compute, data, and specialized labour that determine who can build (Widder et al., 2023). Openness mitigates dependence; it does not abolish it.

A second objection draws on the feasibility evidence itself. If Malagon et al. (2025) show that sovereign-scale training is achievable for middle-income states at costs in the low tens of millions, then the structural pessimism of this article is misplaced. The article has already conceded the empirical point and regards it as genuinely important. But the inference from feasibility to sovereignty does not follow. The modelled feasibility is itself conditional on accelerators procured from a concentrated foreign supply, on the efficiency of hardware designed elsewhere, on remaining within

thresholds set by another state, and on a deliberate decision to forgo the frontier. Each of these conditions is a form of the very dependence the article describes. Feasibility under externally set conditions is precisely what it means to operate within someone else's structure.

A third objection is that the argument counsels passivity — that it risks telling communities not to bother. This would be a serious error, but it is not the article's claim. The position advanced here is emphatically that local training should be pursued and expanded; the achievements documented by Nekoto et al. (2020) and by Te Hiku Media are valuable in themselves and would be valuable even if the larger structural transformation were never achieved. The satellite precedent supports this rather than undermining it: regional broadcasting was worth building even though it did not deliver sovereignty. The argument is not that local training is pointless but that it is incomplete, and that its incompleteness is best addressed not by abandoning it but by pairing it with reform at the levels the structural paradox identifies.

A fourth objection runs in the opposite direction, warning that the very language of "sovereignty" risks legitimizing forms of state control that may themselves be extractive or repressive. "Sovereign AI," as a slogan of national technology policy, can serve the consolidation of state power as readily as it serves communities, and the concern is not hypothetical: a substantial literature documents the procurement of AI systems by states and their agencies for policing and crime prevention, domains in which the people affected are neither customers nor participants but subjects of a capability purchased on their behalf and applied to them (Salehi & Habib Zadeh Khiyaban, 2025). This objection has particular force in the present framing, because the NWICO demand was itself made by states, and its critics argued at the time that a doctrine of national control over information flows would be used to justify censorship as readily as to redress inequality. That criticism was not baseless, and the history should be read as a warning rather than a model. The sovereignty at issue here is not, in the first instance, the sovereignty of states over their populations' data, which can reproduce extraction under a national rather than a corporate flag; it is the authority of communities over the linguistic and cultural resources that are theirs (Kukutai & Taylor, 2016; Carroll et al., 2020; Munn, 2024). The fourth condition specified below — that value return to the communities whose data and labour make AI possible — is what distinguishes community sovereignty from its statist simulacra. The same tension arises within wealthy states as well as between them: research on the deployment of AI in public planning finds practitioners themselves calling for participatory frameworks and accessible tooling rather than the uncritical importation of technocratic paradigms (Sharajsharifi et al., 2025).

A fifth objection targets the historical comparison directly, and it deserves the most careful answer, since Section 3 carries much of the article's weight. Broadcasting and model training, the objection runs, differ in ways that defeat the analogy. A television programme is produced once and transmitted; a model is trained on material of diffuse provenance and then deployed generatively, so the relation between input and output is transformation rather than transmission. The layers are differently arranged: a broadcaster who owned its transponders would have been close to sovereign, whereas a community that owned its accelerators would still depend on frameworks, base models, and the terms of their release. And the satellite era's asymmetries ran between states, in an intergovernmental arena with treaties and allocation procedures, whereas the present ones run between firms and communities, cutting across states rather than lining up with them — a difference Birhane (2020) makes central.

These disanalogies are real, and the argument does not rest on the parallel being exact. What the satellite episode supplies is not a template but a demonstration: that the gap between productive capacity and structural control is a recurrent feature of communication technologies rather than an artifact of this one; that the gap was named and analysed long ago; and that the analytical vocabulary developed for it transfers usefully. If anything, the disanalogies sharpen the warning rather than blunting it. A community acquiring accelerators is further from sovereignty than a broadcaster acquiring a transponder, because the number of

layers over which control would have to be won is larger. And the shift from an intergovernmental arena to a corporate one removes the very forum in which the earlier demand was at least formally heard: there is no UNESCO for base-model licensing.

Engaging these objections does not weaken the central thesis; it locates it more precisely. Local training is necessary, valuable, and more feasible than fatalism allows; it is also, on its own, structurally insufficient; and the response to that insufficiency is neither passivity nor nationalism but a broader, community-centred transformation of the conditions under which AI is built.

11. WHAT GENUINE SOVEREIGNTY WOULD REQUIRE

If local training in the thin sense is insufficient, what would suffice? The literature surveyed here points toward four structural conditions, corresponding to the four layers of Table 2: none of them primarily technical, all of them necessary, and each requiring coordinated political and economic action rather than engineering alone.

The first condition concerns the governance of data. The prevailing default — that human expression appearing online is available for extraction unless specifically protected — would need to be reversed in favour of a presumption that data generated by communities and individuals belong to them and may be used to train models only under terms to which they have knowingly agreed and from which they derive benefit. The data sovereignty tradition (Kukutai & Taylor, 2016), the CARE principles of collective benefit and authority to control (Carroll et al., 2020), and their translation into operational criteria (Munn, 2024) supply the normative architecture for such a reversal, and the broader decolonial turn in data and technology research (Couldry & Mejias, 2023) supplies the analytic case for its urgency. Instruments such as the Kaitiakitanga License show that the architecture can be given contractual form. But the limited uptake of that instrument, and the open questions about enforcing it against parties who never sought access in the first place, indicate how much of the work remains legal and institutional rather than normative.

There is an institutional implication here that the literature on AI sovereignty has largely overlooked, and it follows directly from Section 6. Broadcasters — particularly public service and community broadcasters — possess an unusual combination of the assets these conditions require: substantial archives in languages the web does not hold; public-interest mandates that already oblige them to serve linguistic communities; governance structures with some accountability to those communities; and, in many countries, the only technical staff outside the commercial sector with experience of managing large audiovisual holdings. Te Hiku Media demonstrates that this combination can be converted into community-governed language technology, and the Guinean stations that donated their archives demonstrate a weaker but more replicable version of the same custodial role. Whether other broadcasters can or will do the same is an empirical question this article cannot settle, and there are obvious reasons for caution: state broadcasters are frequently instruments of governments rather than of communities, and the distinction between public service and state control is precisely what the fourth objection warns about. Rights obligations to producers and performers may also constrain what an archive can lawfully release. But the possibility deserves a place on the research and policy agenda, and at present it does not have one.

The second condition concerns computational infrastructure. Closing the compute divide documented by Ahmed and Wahed (2020) cannot be accomplished through technology transfer or capacity-building programmes alone; it requires sustained capital investment in the physical substrate of computation at a scale and over a duration that current efforts do not approach. The feasibility evidence of Malagon et al. (2025) is encouraging in showing that, for non-frontier models, the required investment is not beyond the reach of middle-income states; but it is still substantial, it still depends on hardware procured from a concentrated foreign supply, and it still presupposes a political decision to treat such

infrastructure as a public priority. Proposals for public compute provision — national or regional research infrastructure that allocates computational resources outside the control of dominant firms — follow directly from the compute-divide analysis, and represent one of the few mechanisms capable of addressing the constraint at the level at which it operates. The regional cooperation that produced shared satellite capacity in the Arab world and elsewhere is a precedent worth examining here, including for its failures.

The third condition concerns frameworks and base models. Genuine sovereignty would require not merely the existence of open frameworks and open multilingual base models, but their development and governance through processes that are themselves international and participatory, and on terms that cannot be revoked at the discretion of a single firm. Consortium-built models such as BLOOM (BigScience Workshop et al., 2023) point toward this possibility. But two qualifications, drawn from Section 8, must be attached to any such programme. Openness must be specified rather than asserted, since it is a gradient across many dimensions and the label alone guarantees very little (Liesenfeld & Dingemanse, 2024); and openness must be recognized as insufficient on its own, since licensing terms do not redistribute the compute, data, and labour that determine who can build in the first place (Widder et al., 2023).

The fourth condition concerns the distribution of value. At present, the value generated by AI development — economic and epistemic alike — accrues overwhelmingly to the firms that build the models, their investors, and a relatively small technical workforce, while the communities whose data and labour make the systems possible receive little. Genuine sovereignty would require mechanisms of benefit-sharing and reciprocity: arrangements ensuring that value created from a community's data and labour returns to that community, whether through compensation, shared ownership, technology-sharing obligations, or other institutional means. Such mechanisms are not merely aspirational. The principle of collective benefit is the first of the CARE principles (Carroll et al., 2020), and the Kaitiakitanga License operationalizes it as a condition of access rather than a statement of intent. That so few comparable instruments exist, and that their reach remains largely untested, is itself the measure of how far this condition is from being met. This is not merely a matter of fairness, though it is that; it is a matter of sustainability, since a system that expects communities to contribute their linguistic and cultural inheritance while capturing the resulting value elsewhere is neither just nor stable. The decolonial analysis of AI (Mohamed, Png, & Isaac, 2020), with its diagnosis of algorithmic dispossession, makes clear that the distribution of value is a central axis of the problem rather than a downstream detail.

These four conditions are demanding, and none can be secured by technical means alone. They are political and economic transformations, bearing on ownership, power, and the allocation of resources. Local training is a component of the larger reorganization they describe, and an important one; but it is a component, not a substitute, and treating it as the whole is precisely the error the structural paradox exposes.

12. CONCLUSION

The movement toward local and community-controlled AI training is a serious and constructive response to a genuine asymmetry in the global development of artificial intelligence. The initiatives it has produced — the participatory, continent-spanning work of networks such as Masakhane, the corpora and benchmarks assembled for the languages of South Asia and the Americas, the community-governed speech resources built by the Māori broadcaster Te Hiku Media — have achieved real improvements in capability for long-neglected languages, have built research capacity and institutions where little existed, and have asserted, in practice, the principle that communities should govern the technologies that mediate their linguistic and cultural lives (Nekoto et al., 2020; Birhane, 2020; Kakwani et al., 2020; Mager et al., 2021).

What critical analysis reveals is that local training, pursued in isolation, encounters limits set by parts of the system that lie beyond local control. The data on which models are built remain governed

by defaults that presume extraction, and where the richest holdings sit in broadcast archives they remain governed by institutions whose accountability to language communities varies enormously. The hardware and infrastructure on which training depends remain concentrated and subject to foreign policy. The frameworks and base models that mediate development remain shaped and provisioned by a small number of actors, under terms whose openness is frequently more marketed than specified. And the value that AI generates continues to flow disproportionately to those actors rather than to the communities whose data and labour make it possible. Recent feasibility evidence (Malagon et al., 2025) usefully complicates any claim that compute renders local training simply impossible — for non-frontier models, sovereign-scale training is more attainable than fatalism suggests, and export controls are not the binding constraint they are often assumed to be — but it equally underscores the depth of the dependence that persists even where training is feasible. Feasibility is not sovereignty.

That conclusion should not surprise anyone who has studied broadcasting. The demand for communicative sovereignty has been made before, in terms very close to these, and the answer it received was neither victory nor defeat but asymmetrical interdependence: production capacity genuinely won, structural control genuinely withheld. The field of communication has had a name for this condition since 1991 and a diagnosis of its causes since 1980. The contribution this article claims is to bring that inheritance to bear on a debate that has been conducted almost entirely without it, and to specify where, in the case of model training, the layers of control now sit.

Genuine linguistic and computational sovereignty would therefore require more than the capacity to train models on local soil. It would require reversing the default governance of data — which, for spoken language, means confronting the question of who holds the archives and on what terms they release them; building and governing computational infrastructure as a public good; provisioning frameworks and base models through durable international commons; and constructing mechanisms that return the value of AI development to the communities that make it possible. These are political and economic tasks, not technical ones. The path forward does not lie in a choice between local training and structural reform. It lies in pursuing both at once, with a clear understanding of what each can and cannot accomplish. Pursued that way, local training is an indispensable part of a larger project of justice in artificial intelligence; mistaken for the whole of that project, it risks delivering the form of sovereignty while leaving its substance in other hands — which is, precisely, what happened last time.

REFERENCES

Ahmed, N., & Wahed, M. (2020). *The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research* (arXiv:2010.15581). arXiv. https://doi.org/10.48550/arXiv.2010.15581

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science, 379*(6635), 884–886. https://doi.org/10.1126/science.ade2420

BigScience Workshop, Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., … Wolf, T. (2023). *BLOOM: A 176B-parameter open-access multilingual language model* (arXiv:2211.05100). arXiv. https://doi.org/10.48550/arXiv.2211.05100

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal, 19*(1), 43. https://doi.org/10.5334/dsj-2020-043

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

Couldry, N., & Mejias, U. A. (2023). The decolonial turn in data and technology research: What is at stake and where is it heading? *Information, Communication & Society, 26*(4), 786–802. https://doi.org/10.1080/1369118X.2021.1986102

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dariush, B., Dastyar, M., & Dastyar, F. (2025). Environmental intelligence and globalization: Redefining educational paradigms through an Iranian perspective. *Journal of Cyberspace Studies, 9*(2), 369–386. https://doi.org/10.22059/jcss.2025.390799.1130

Dastyar, M., Dariush, B., & Dastyar, F. (2023). Virtual education, a complementary element of the puzzle of holistic education in the relationship between humans and environment for Iranian citizens. *Journal of Cyberspace Studies, 7*(2), 237–252. https://doi.org/10.22059/JCSS.2023.356702.1088

Doumbouya, M., Einstein, L., & Piech, C. (2021). Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(17), 14757–14765. https://doi.org/10.1609/aaai.v35i17.17733

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Habib Zadeh Khiyaban, S., & Sabbar, S. (2026). Artificial intelligence in banking: Opportunities, risks, and ethical imperatives. *Code, Cognition and Society, 1*(2), 1–12. https://doi.org/10.22034/ccsr.2025.548684.1017

Hao, K. (2022, April 22). A new vision of artificial intelligence for the people. *MIT Technology Review*. https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/

International Commission for the Study of Communication Problems. (1980). *Many voices, one world: Towards a new, more just and more efficient world information and communication order*. UNESCO.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Joynes, J., Dwivedy, A. K., & Kumar, A. (2026). NWICO 2.0: A decolonial framework for rethinking global digital communication. *Annals of the International Communication Association*, wlag025. https://doi.org/10.1093/anncom/wlag025

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4948–4961). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.445

Kukutai, T., & Taylor, J. (Eds.). (2016). *Indigenous data sovereignty: Toward an agenda*. ANU Press.

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1774–1787). Association for Computing Machinery. https://doi.org/10.1145/3630106.3659005

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., … Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for Indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas* (pp. 202–217). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.23

Malagon, S., Ulloa Ruiz, M. A., Sandoval Plaza, T. E., Rosario Bolívar, G. R., García Mesa, V., & Alvarado Morales, I. (2025). *The feasibility of training sovereign language models in the Global South: A study of Brazil and Mexico* (arXiv:2510.19801). arXiv. https://doi.org/10.48550/arXiv.2510.19801

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Munn, L. (2024). The five tests: Designing and evaluating AI according to indigenous Māori principles. *AI & Society, 39*(4), 1673–1681. https://doi.org/10.1007/s00146-023-01636-x

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., … Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144–2160). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Sabbar, S., & Habib Zadeh Khiyaban, S. (2023). Algorithms of displacement: Emotional and rhetorical responses to AI-driven job loss in digital public discourse. *International Journal of Advanced Multidisciplinary*

Research and Studies, 3(4), 1324–1331. https://doi.org/10.62225/2583049X.2023.3.4.5012

Sakr, N. (2001). *Satellite realms: Transnational television, globalization and the Middle East*. I.B. Tauris.

Salehi, K., & Habib Zadeh Khiyaban, S. (2025). AI and crime prevention in the academic literature: An integrative review of AI applications in crime prevention. *Code, Cognition and Society, 1*(1), 164–177. https://doi.org/10.22034/ccsr.2025.546552.1016

Sharajsharifi, M., Sharifi Poor Bgheshmi, M. S., & Aeini, B. (2025). Artificial intelligence for equitable urban housing: Insights from U.S. sustainable planning experts. *Code, Cognition and Society, 1*(1), 141–163. https://doi.org/10.22034/ccsr.2025.531875.1011

Straubhaar, J. D. (1991). Beyond media imperialism: Asymmetrical interdependence and cultural proximity. *Critical Studies in Mass Communication, 8*(1), 39–59. https://doi.org/10.1080/15295039109366779

Te Hiku Media. (2022). *Data sovereignty and the Kaitiakitanga License*. https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license

Widder, D. G., West, S., & Whittaker, M. (2023). *Open (for business): Big Tech, concentrated power, and the political economy of open AI*. SSRN. https://doi.org/10.2139/ssrn.4543807