Extractive Patterns and Equitable Pathways: A Landscape Analysis of AI Data Production Through Pipeline and Historical Perspectives

ANONYMOUS AUTHOR(S)

Artificial intelligence depends on data production processes that transform human knowledge into computational resources. Many approaches reproduce extractive dynamics rooted in histories of appropriation and dispossession. We conduct a multivocal landscape analysis of 350 sources on the relationship between AI systems, data production, and impacts on Indigenous, underrepresented, and underserved communities worldwide. Our synthesis identifies twelve extractive patterns and eleven less-extractive alternatives across the AI development pipeline. The literature exhibits an action gap: 257 sources diagnose extractive practices or advance normative principles, yet only 93 document concrete alternative practices. We contribute (1) a conceptual shift from data collection to data production that foregrounds sociotechnical complexity and power dynamics; (2) empirical mapping of sources across pipeline stages, historical eras, and geographic contexts; (3) an interpretive synthesis methodology that bridges critical theory and historical foundations with technical analysis; and (4) an open corpus and interactive tools to support further inquiry.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: AI data production, extractive practices, underserved communities, Indigenous data sovereignty, pipeline analysis, data practices, multilingual, cultural context, data collection

ACM Reference Format:

Anonymous Author(s). 2018. Extractive Patterns and Equitable Pathways: A Landscape Analysis of AI Data Production Through Pipeline and Historical Perspectives. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 44 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Sociotechnical perspectives that treat technology as inseparable from its social and political conditions are widespread across disciplines. Human-computer interaction (HCI) works offer increasingly sophisticated frameworks for this analysis [39, 51]. A broad body of works ground artificial intelligence (AI) development in longer global histories of appropriation, dispossession, and unequal exchange, framing data practices as extractive [87, 206] and colonialist [17, 27, 130]. Indigenous [114, 161], feminist [99], postcolonial [215], decolonial [21, 135], and other approaches show how AI reproduces inequities and where community-led alternatives emerge, or could.

Extensive studies establish cultural bias in LLMs as a pervasive problem, one that reflects how technology embeds both its creators' perspectives and the human knowledge it's built with [140, 162]. Researchers have responded in the following complementary directions. Provenance studies move backwards from training and test corpora. Early interventions [56, 83] and ongoing initiatives [5] trace origins and register whose perspectives dominate or are absent. Cultural alignment studies work outwards from model behavior, developing evaluation datasets and auditing methods that assess how systems reflect diverse values across different cultural contexts [13, 14, 96, 171]. Dataset

Manuscript submitted to ACM

creation initiatives work forwards from local priorities, building training resources with attention to and engagement with cultural and linguistic contexts. Efforts include multilingual training corpora developed through international collaboration [109], independent data resources in Global Majority regions [31], and accessibility-focused collections [2].

This landscape analysis brings diverse streams together, from critical foundations to technical approaches, to examine AI data production and its impacts on Indigenous, underrepresented, and underserved communities worldwide. Here we define *data production* as encompassing all processes that render human knowledge, practices, and infrastructures into computational form.[1] The production frame moves beyond narrow notions of *data collection* as either a routine methodological step or the accumulation of neutral technical artifacts. The shift from collection to production opens a more situated view of how data practices unfold across the AI data production pipeline and brings into focus extractive patterns, emerging equitable pathways, and the uneven dynamics that shape them.

AI adoption is rising globally. Global Majority regions both drive this growth and disproportionately supply the data enrichment crucial to those systems [1, 211]. Training and test data provenance is heavily concentrated in North America and Europe [121]. A structural asymmetry shapes research infrastructure, too. Studies demonstrate that computational venues exhibit Western-centric biases and devalue non-Western epistemologies in sociotechnical applications [11, 187, 200]. This multidimensional situation requires methods that synthesize multiple dimensions simultaneously.

Our landscape analysis maps extractive patterns and less-extractive alternatives across AI development pipeline stages and historical eras of data practices. We offer two primary contributions to HCI and critical computing research, supported by methodological and resource contributions:

Primary contributions:

(1) Conceptual reframing: We shift the unit of analysis from data collection to data production, foregrounding sociotechnical complexity and power dynamics in how human knowledge becomes computational resources.
(2) Empirical synthesis: We systematically code and analyze 350 sources to synthesize thirteen extractive patterns and twelve less-extractive pathways. The literature exhibits a critical action gap where critique and principles outweigh documented practices nearly three-to-one.

Supporting contributions:

(1) Methodological approach: An interpretive synthesis methodology that bridges critical theory and historical foundations with technical analysis across diverse evidence types.
(2) Open resources: A dataset of 350 sources with bibliographic metadata and our qualitative codes, plus interactive query environments. Available at https://chifod2025.github.io/ai-data-production-landscape/.

1.1 Key terms and definitions

1.1.1 Artificial intelligence. We define AI as data-intensive computational systems trained on human-produced content, behavior, and labor to generate outputs including predictions, classifications, recommendations, or content [145]. Examples include machine learning (ML), natural language processing (NLP), computer vision, and LLMs. Systems that generate synthetic data remain derivative of prior human production. We use the term in this technical sense and acknowledge variation in autonomy, adaptiveness, and context of deployment.

[1] We share with Miceli & Posada [132] an emphasis on "production" to foreground relations of power and knowledge in data and labor, which echoes the "assemblage" approach of Kitchin et al. [97], also rooted in Foucauldian critique.

1.1.2 Extractive. We define extractive as non-sustainable resource production that exploits individuals and communities by undermining meaningful consent, fair compensation, and/or tangible benefit [61]. Extractive practices target or exclude historically oppressed populations and generate value primarily for downstream refiners, marketers, and integrators rather than proximally situated resource workers. Indigenous scholar Leanne Betasamosake Simpson describes the stakes: *"The act of extraction removes all of the relationships that give whatever is being extracted meaning. Extracting is taking. Actually, extracting is stealing—it is taking without consent, without thought, care or even knowledge of the impacts that extraction has on the other living things in that environment" [100]*

1.1.3 Communities and populations. We use underserved to describe communities lacking adequate infrastructural, institutional, or economic support, and underrepresented to indicate groups whose knowledge, languages, or perspectives are numerically absent or devalued in AI research and development [115, 189]. We use the umbrella category "Indigenous," which "enables historically and geographically separated peoples to recognize each other and their common plight, and to collaborate towards a better future" [175]. We avoid "marginalized" in the adjectival form to emphasize agency and resistance rather than positioning communities as passive victims. We use Global Majority to emphasize that most of the world's population lies outside Euro-American contexts. Our chosen terms underscore structural asymmetries in power and resource distribution rather than deficits within communities themselves [201].

1.2 Positionality

The authors come from both industry and academia and are based in Europe, South Africa, the UK, and the U.S. We are interdisciplinary, with backgrounds in technology, social science, and the humanities. Across these domains we share a history of publishing in advocacy for social justice, from computing and data practices to broader cultural questions of power, exclusion, and representation. The foundational research for this paper was supported by an industry partner redacted for anonymity. The writing and publication were prepared independently.

Our diverse institutional roles and geographic locations shape how we approach the politics of AI data production. Industry experience grounds our analysis in product realities and data practices at scale. Academic and humanistic experience bring theoretical and cultural perspectives from HCI, Science, Technology, and Society (STS), decolonial thought, and critical studies of culture, religion, and gender. We situate our work within traditions that emphasize accountability, data justice, and the interrogation of extractive practices. We recognize the limits of our own perspectives and do not seek to speak for all communities.

2 Related Work

2.1 Critical traditions on extraction and justice

Foundational works from theoretical, historical, and community traditions establish frameworks for studying power in knowledge production. Theories of epistemic violence and injustice [68, 194] and "situated knowledges" [77] interrogate how knowledge systems encode relations of domination. Historical analyses of colonial resistance show alternative epistemologies and organizing strategies [89].

Black feminist theory articulates intersectional approaches to structural power, from early collective statements [42] to analyses of interlocking systems of oppression [43]. Gender and queer theory establish frameworks for analyzing the production of normativity [35], binary logics [186], and classificatory power [41]. Indigenous studies center community sovereignty and relational ethics [22, 54, 113] and provide frameworks for decolonizing knowledge production in

research [192] and AI data practices [34]. Critical data studies crystallize a complementary set of concerns for the digital context, with a focus on datafication, surveillance, and governance [98].

Foundational works attune us to centuries of extractive patterns, resistance, and knowledge-making. They are essential for understanding present and future technological worlds. Here, critical works anchor the conceptual vocabulary of extraction and justice in the context of the global AI data production ecosystem.

2.2 Evolution of AI data production practices

Since the advent of machine learning, there has been a constant need for data. Over time, how that data was produced has undergone transformations beyond dataset sizes. These changes include how data is produced and who performs the work [56]. As demands for larger models have intensified, practices have shifted from small, carefully curated corpora, to large benchmark datasets assembled through crowdsourcing, to massive, automated web-scraped collections supported by industrial-scale annotation.

Era 1. Early curated datasets were small, domain-specific, and selected by experts. A canonical example is MNIST, a dataset of handwritten digits drawn from U.S. postal codes [110]. Choices about inclusion and categorization reflected institutional knowledge and disciplinary priorities.

Era 2. Large curated datasets expanded scale through crowdsourced annotation, exemplified by ImageNet [55] and MS COCO [118]. This era accelerated deep learning [52, 75] but shifted labor from domain experts to distributed workers, often in Global Majority regions, and emphasized performance gains over contextual fit.

Era 3. Contemporary data production diverges into two parallel approaches. Massive, largely uncurated web-scraped corpora such as Common Crawl, C4 [58, 172], LAION [183, 184], Refined Web [160], and ClueWeb22 [153]) are assembled through automated scraping at an unprecedented scale. Such production efforts shape and are shaped by competitive foundation model development [29, 199]. Alongside and often in response, smaller, highly curated datasets emerged, produced through participatory methods and community partnerships. Examples include ROOTS [109], Masakhane's African language collections [142], and Cohere's multilingual Aya Dataset [191].

Dataset hosting and governance practices have shifted over time as well: from freely downloadable units like MNIST, to single-location storage on cloud services (e.g., AWS, HuggingFace), to URL-indexed collections like LAION that disclaim responsibility for original sources, and to emerging federated "data spaces" designed to support locally owned infrastructures and community governance [84]. Proprietary datasets are closed, and open-source alternatives range from massive scrapes to carefully stewarded community collections; each modality comes with distinct risks and obligations [116, 215].

The three eras feature distinct technical capabilities, institutional arrangements, and methodologies that enabled extractive patterns to scale industrially. The current era hosts both the most expansive extractive practices and the most developed community-controlled frameworks. The future of AI data production is not determined.

2.3 Review methodologies and synthesis challenges

2.3.1 Substantive review traditions in AI data production. Four review traditions establish precedents for organizing AI data production literature, each with distinct strengths and scope limitations. Sociotechnical and data justice reviews foreground historical and structural conditions of AI development and deployment. Authors trace patterns of engagement and representation [29, 189] and show how institutional values shape data infrastructures [112, 213]. The strength of this tradition lies in connecting AI systems to governance and power. Technical syntheses establish taxonomies of risks and interventions and supply shared vocabularies for fairness, bias, and privacy [15, 38]. Critical

surveys locate these vocabularies within labor, culture, and lived experience [33, 158]). Such works clarify technical levers and situate them in social context. Genealogical accounts map dataset cultures and document how scale, labor, and provenance shape research practices [56, 83]. This tradition positions documentation and transparency as cornerstones of accountability. Participatory and community-collaborative reviews specify criteria for trust, role clarity, and benefit sharing [45, 53, 65]. Their thematic depth is a strength, even when insights remain tied to particular projects and domains.

Our work builds on these traditions through multidimensional synthesis. We examine extractive patterns and less-extractive alternatives across historical eras, pipeline stages, and geographic regions. This approach bridges technical and social analysis to identify implementation gaps and surface ecosystem-level alternatives emerging from community-led initiatives worldwide.

2.3.2 Review methodology challenges and innovations. Review methodologies in HCI and computing establish protocols for organizing complex literatures and clarify scope conditions that shape what reviews can accomplish. Systematic and scoping reviews establish protocols for organizing evidence and validating findings across bounded domains. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) strengthens transparency and reproducibility [154]. These frameworks provide value across many domains, including HCI. They originated in medical sciences and spread to software engineering, two domains where research questions tend to align with well-defined boundaries and established publication ecosystems [16, 139]. However, such strictures reach limits of adaptability in interdisciplinary domains where evidence types vary and epistemological frameworks conflict.

Alternative synthesis methodologies address these limitations through different approaches. Meta-ethnography [143] translates qualitative studies across contexts to build higher-order interpretations, and realist synthesis [159] draws out "what works, for whom, in what circumstance" across complex interventions. Interpretive methods accommodate diverse evidence types while preserving meaning and context.

In HCI specifically, methodological challenges have prompted calls for reform. Rogers et al. [174] identify persistent ambiguity in computing survey purposes and practices, and issue guidance to help researchers undersatnd and clarify their reporting approach. Fok et al. [67] show that narrative reviews face various update challenges, including empirical, structural, and interpretive, and point out the difficulty of sustaining updating amid busy working lives. They call for methods that combine systematic transparency with interpretive expertise.

Our approach responds to these calls with a multivocal design [71] that integrates white and grey literature, including community outputs, technical reports, and organizational documentation. We also draw on critical theory and historical work. Decolonizing methodologies guide how we name exclusions [22, 192]. Indigenous scholarship provides epistemological foundations that foreground sovereignty, relationality, and accountability [114].

3 Methodology

3.1 Corpus construction

We assembled 350 sources through a multivocal search strategy designed to surface perspectives often excluded from academic venues. Our approach accommodated the interdisciplinary nature of AI data production inquiry while including Indigenous, underrepresented, and underserved voices.

Inclusion strategy. A source was included if it substantively engaged at least one vertex of the triangular relationship between AI systems (A), data production (D), and community impacts (C). We call this the A/D/C triangle (see Figure 1).

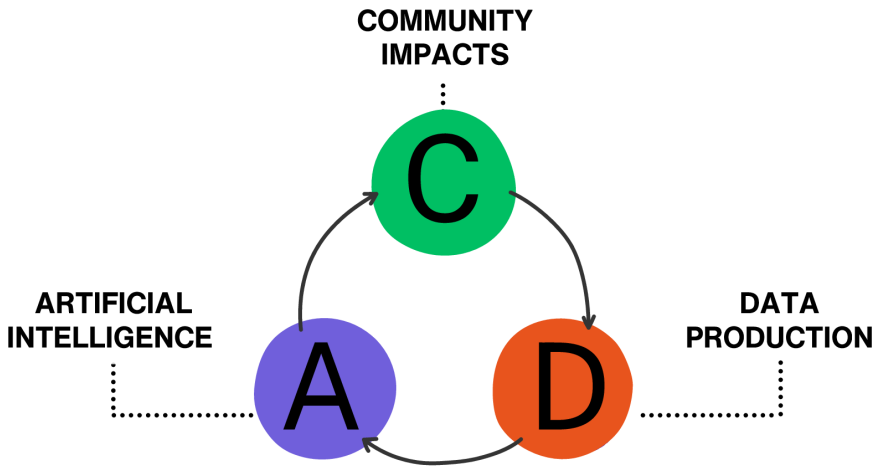

Fig. 1. A/D/C analytical framework for source analysis. Sources were coded based on substantive engagement with Community Impacts (C), Artificial Intelligence (A), and/or Data Production (D). Arrows indicate the interconnected relationships between dimensions that inform AI data production practices.

This approach allowed us to capture the interdisciplinary nature of AI data production inquiry across technical, social, and community domains.

Search process. Database searches across eight venues (175 sources, 50%) employed a two-phase strategy targeting foundational AI data terminology and community impacts, then expanding to extractive/less-extractive frameworks and community-centered methods (see B). Interpretive search extensions (175 sources, 50%) actively sought sources through network recommendations, citation snowballing from highly relevant sources, and hand-searching key venues like FAccT and CHI. Grey literature (92 sources, 26%) included organizational reports, policy documents, community statements, and technical documentation from policy initiatives, and NGO/think tank publications.

3.2 Coding framework

3.2.1 Pipeline stages. We coded each source to stages of a simplified AI development pipeline (Figure 2; see C for complete coding definitions): *Problem Understanding & Formulation* (institutional prioritization, funding decisions, and product conception), *ML System Design and Development* (data selection and enrichment, model architecture choices, and training processes), and *Deployment & Impact* (product testing, launch, and post-deployment effects) (Martin, 2020).

Although presented linearly, AI development involves overlapping, cyclical processes where deployment failures can reshape problem formulation, evaluation metrics redefine training objectives, and community pushback alter institutional priorities [197]. The interconnected nature of these processes means that problems at one stage "cascade" through the pipeline, compounding downstream issues and harms [179].

3.2.2 Historical eras. We distinguished three eras of data production, per discussion in 2.2: Era 1 (expert-curated datasets, pre-2009), Era 2 (crowdsourced benchmarks, 2009–2017), and Era 3 (web-scraped and foundation models, 2017–present). Multi-era sources were coded accordingly. No sources were coded exclusively as Era 1.

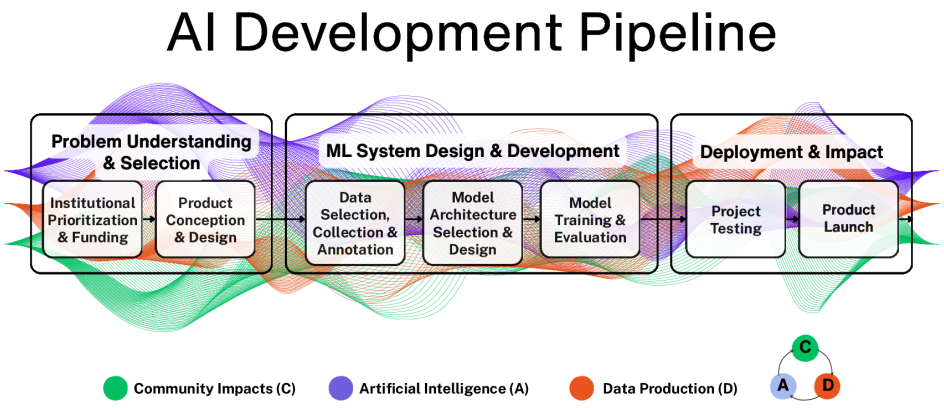

Fig. 2. AI Development Pipeline with interconnected analytical dimensions. Three sine waves represent Community Impacts (green), Artificial Intelligence (purple), and Data Production (orange) flowing continuously through all pipeline stages, illustrating how these dimensions are present throughout the development process rather than confined to specific stages. AI development pipeline derived from Martin [129].

3.2.3 Orientations. Three orientations captured how sources positioned data production:

(1) **Extractive:** Sources documenting, analyzing, or critiquing practices that undermine meaningful consent, fair compensation, or tangible community benefit
(2) **Less-extractive principles:** Sources advancing normative approaches including theoretical frameworks, policy recommendations, conceptual models, and advocacy statements that promote stewardship, sovereignty, and accountability
(3) **Less-extractive practices:** Sources documenting operationalized examples of community-led, participatory, or sovereignty-based initiatives

3.2.4 Contextual dimensions. Contextual dimensions coding assessed whether sources substantively engaged AI systems (A), data production practices (D), and community impacts (C). Substantive engagement meant analysis beyond a passing mention, constituting either a primary focus or a significant analytical component that informed relationships across vertices, even if not made explicit.

Single-vertex codes applied when sources addressed only community impacts without substantive engagement with AI systems or data production practices (C and D only), for example, governance frameworks discussing AI regulation without technical detail or community advocacy statements documenting harms without analyzing underlying AI systems or data processes. Multi-vertex codes captured intersections: dataset papers linking collection methods to bias outcomes (AD), advocacy statements connecting AI systems to social harms (AC), or labor studies examining both working conditions and downstream model effects (DC).

Sources addressing all three vertices of the triangle (ADC) formed the analytical core of our synthesis. Participatory, community-engaged dataset initiatives (e.g., [94, 207]) are a prime example: they create corpora for use in AI contexts and address specific community needs.

3.2.5 Coding procedures. Inter-coder reliability was established through collaborative consensus across multiple validation rounds (see B). The team collectively identified foundational sources, after which the first author completed the full corpus coding with ongoing oversight and consensus-building.

Each source received one orientation code (extractive, less-extractive principle, or less-extractive practice). Within that orientation, sources could be coded to more than one specific pattern or pathway if their contributions substantively spanned multiple mechanisms. For instance, a source analyzing both data labor exploitation and biased preprocessing was coded to patterns 9 #8 and #9.

Note: Counts under annotation headers reflect coding tags, not unique sources. Because a single source may generate multiple tags, total pattern and pathway tallies (n=416) exceed the number of unique sources in the corpus (n=350). Each source, however, carries only one orientation code.

3.3 Synthesis procedure

We synthesized findings by first mapping sources to pipeline stages and sub-stages to locate where dynamics occurred, then clustering by domain tags (data labor, data practices, ethics frameworks, community impacts, critical theory/historical background) to surface recurring concerns within and across stages. Triangle coverage (A/D/C) served as a cross-cutting lens to identify patterns of engagement: whether sources addressed single vertices, paired combinations, or integrated all three dimensions.

Orientation coding required resolving sources that documented both harms and alternatives. Codes reflected the primary analytical purpose the source served in our interpretation for this inquiry. For example, many works centered on harmful dynamics were coded as extractive even if they proposed remedies, and some advancing frameworks or practices were coded as less-extractive even if they acknowledged extraction. We distinguished orientations through close reading and discussion.

From this coded corpus, we identified extractive patterns as recurring mechanisms and less-extractive, more equitable pathways as documented alternatives. Pathways had more examples since they often appeared as specific initiatives rather than structural diagnoses. Sources contributing conceptual, historical, or framing insights without mapping directly onto patterns or pathways were coded as Other/NA. Era mapping situated contributions temporally to characterize mechanisms and alternatives recurring across stages, domains, and eras.

3.4 Limitations and reflexivity

Manual coding and collaborative discussion constrained this methodology. The labor-intensive approach enabled deep engagement with sources but limited corpus size and prevented more expansive coverage. We recognize that subjectivity inevitably shapes our interpretations, though transparent documentation and multi-researcher validation helped mitigate this risk. Connecting multiple disciplinary traditions, historical eras, and global contexts proved challenging, creating inherent "translation needs" across distinct vocabularies and epistemological frameworks.

The sourcing strategy privileged networks in Africa and global Indigenous movements, yielding detailed coverage of those ecosystems. Parallel developments in Arabic, Southeast Asian, and Latin American contexts appear less frequently, not because such initiatives were absent, but because they circulated in networks less accessible to our inquiry. English-language search restrictions further reinforced Western-centric representation, despite active efforts to surface materials by Indigenous and Global Majority authors through academic venues, organizational reports, and community outputs. Scholarly and community infrastructures condition what becomes visible in review corpora, creating this unevenness despite our efforts. Our positioning across industry and academia exposes us to the same

structural biases we identify in the corpus. We recognize that our analysis risks reproducing these dynamics even as we aim to surface alternatives. We present these patterns as a first step toward change.

Beyond sourcing limitations, the corpus composition reveals substantive gaps. Orientation codes demonstrate a marked imbalance: 257 sources analyzed extractive practices or advanced less-extractive principles, compared with 93 documenting operational practices. In short, nearly three sources describe extractive practices or articulate normative principles for every source detailing concrete less-extractive practices. This action gap constitutes a key finding as much as a limitation of our approach.

The temporal cut-off of January 2025 inevitably excluded works and community-led initiatives not yet discovered or formally documented. Extremely dynamic developments in many relevant fields mean academic review processes lag behind emerging discourse and community practices.

4 Findings

From the 350-source corpus, we identify twelve extractive patterns and eleven less-extractive pathways across the AI development pipeline. Each annotation in the sections below define the pattern or pathway, illustrates it with selected examples, and provides a brief note on historical evolution. Each annotation is accompanied by counts of corpus sources coded to that pattern or pathway, providing a sense of relative representation across the review. For less-extractive pathways, we emphasize concrete practices over normative principles to highlight actionable alternatives. The annotations are not exhaustive but serve as scaffolding to motivate further inquiry.

4.1 Mapping extractive practices across the pipeline

Table 1 lists the twelve extractive patterns identified across the pipeline and serves as a reference point for the annotations that follow. Counts represent coding tags rather than unique sources, since individual works may map to multiple mechanisms.

4.1.1 Problem Understanding and Formulation.

1. **Excluding underrepresented groups from decision-making**
   Corpus tags: 14 (3% of all tags)
   Problem formulation functions as data production: it determines whose knowledge, needs, and contexts are transformed into computational form. Exclusion from AI governance and agenda-setting concentrates power in economically developed countries and wealthy corporations rather than diverse communities [167]. Communities may be over-represented in datasets yet excluded from governance frameworks, as with Aboriginal and Torres Strait Islander Peoples under Australia's 2022 Data Availability and Transparency Act, which permits data sharing without Indigenous sovereignty provisions despite UN Declaration commitments [37]. Governance gaps translate to Global Majority regions and communities being positioned as data sources and prevented from meaningfully influencing problem selection and resource allocation [19, 150]. Local concerns often go unrecognized as legitimate technical problems, reinforcing patterns where what counts as relevant data reflects Northern institutional priorities rather than global community needs which are diverse and heterogeneous [135, 165]. Weak advocacy structures in global AI governance perpetuate this concentration, giving rise to feedback loops where agenda-setting power shapes data production practices that further marginalize underrepresented perspectives.

Table 1. Twelve extractive data production patterns identified from corpus analysis across the AI development pipeline

Pipeline stage	Pipeline sub-stage	Extractive practices
Problem understanding & formulation	Institutional prioritization & funding Product conception & design	1. Excluding underrepresented groups from decision-making
ML system design & development	Data selection, collection & annotation	2. Collecting vast amounts of data to train AI systems 3. Reproducing biases through synthetic data generation 4. Scraping or repurposing sensitive data
	Model architecture selection & design	5. Prioritizing data wants over community needs 6. Soliciting data without reciprocal benefits
	Model training & evaluation	7. Exploitative and invisible data labor 8. Biased pre-processing and category erasure 9. Keeping communities in the dark through opaque data practices
Deployment & impact	Product testing Product launch	10. Ethics dumping in less-regulated contexts 11. Deploying AI systems trained without local, contextual data 12. Western-centric research infrastructures

Era view. In E1, expert curation limited decision power. In E2, crowdsourcing broadened participation numerically but not governance. In E3, foundation-model agendas are concentrated in a few firms despite global reach.

4.1.2 *ML System Design and Development.*

2. **Collecting vast amounts of data to train AI systems**

Corpus tags: 30 (7% of all tags)

AI scales data aggregation beyond existing oversight capacity, intensifying governance failures through massive collection that outpaces regulatory frameworks. Large-scale data collection treats communities and individuals as raw material for AI systems, and their digital traces as extractable resources [48, 182]. Sources assess legal frameworks as inadequate to prevent these violations [95, 158]. Collection practices may disclose personal information through text scraping, moreover, enable profiling through undisclosed voice harvesting, and may absorb explicit content and racist material from web sources without filtering or acknowledgment [30, 56, 59, 86]. Scraping from underrepresented and underserved populations intensifies the risks when local contexts and vulnerabilities are systematically ignored (see 4 and 12 below). AI scales data aggregation beyond existing oversight capacity, intensifying governance failures through massive collection that outpaces regulatory frameworks. Large-scale data collection treats communities and individuals as raw material for AI systems, and their digital traces as extractable resources [48, 182]. Sources assess legal frameworks as inadequate to prevent these violations [95, 158]. Collection practices may disclose personal information through text scraping,

moreover, enable profiling through undisclosed voice harvesting, and may absorb explicit content and racist material from web sources without filtering or acknowledgment.

Era view. In E1, small curated sets made inclusion choices explicit. In E2, benchmark scale encouraged indiscriminate collection. In E3, web-scale scraping treats nearly all digital traces as potential data.

3. **Reproducing biases through synthetic data generation**

Corpus tags: 8 (2% of all tags)

Synthetic data production replaces human-derived knowledge with algorithmically generated substitutes. Organizations generate fabricated medical records, artificial images, and simulated text to circumvent data scarcity constraints and consent requirements [111]. Synthetic datasets embed the representational failures of their training sources yet sever accountability ties to original communities, creating false confidence in diversity and displacing rather than addressing ethical concerns [78, 147, 216]. Facial recognition tests with synthetic faces create a sense of false diversity that benefits developers while leaving real communities more vulnerable to harm [214]. Quality validation across cultural contexts remains persistently limited, as practitioners struggle to accurately depict communities absent from source datasets [92]. Research applications increasingly propose replacing human participants with AI surrogates, removing direct engagement with communities while maintaining claims to representation [10].

Era view. Synthetic data was not central in E1 or E2. In E3, synthetic data emerges as a dominant practice, reproducing bias while weakening accountability ties.

4. **Scraping or repurposing sensitive data**

Corpus tags: 5 (1% of all tags)

Data production that treats culturally specific or sacred materials as generic training inputs disregards the meanings and vulnerabilities attached to them. Religious texts have been absorbed into NLP datasets without acknowledgment of cultural significance or colonial collection histories, reduced to "just data" [36, 82]. Contextual language use and spiritual significance are overlooked when missionary and colonial archives are ingested as convenient parallel corpora for scaling machine translation systems [12, 47]. Repurposing missionary-led translations for commercial AI systems reproduces historical divisions and extracts community knowledge without consent [125], as in cases of Indigenous oral stories being scraped and placed on commercial servers without permission [91]. *Era view.* In E1, missionary and linguistic archives supplied corpora. In E2, religious and cultural texts became convenient benchmark sources. In E3, foundation models fold sacred and culturally embedded materials into large datasets, often without context.

5. **Prioritizing data wants over community needs**

Corpus tags: 11 (3% of all tags)

AI data production often entails a prioritization of institutional and commercial agendas over community needs, regardless of collection scale. Sources describe data extraction driven by commercial imperatives [177] as a practice devoid of consideration for community contexts or agency [104, 133]. The wants-over-needs dynamic reflects deeper assumptions about whose knowledge forms count as legitimate data sources. The longstanding textual orientation in research privileges written texts as default data forms, which marginalizes oral traditions and creates potentially harmful mismatches—especially when systems trained on written corpora are deployed to communities whose languages are primarily oral [25, 192]. Institutional wants can harm the communities

they claim to serve in other ways too: data collection intended to address inequality can generate harm by placing vulnerable communities at risk through the act of collecting data itself [99].

Era view. In E1, expert projects privileged written corpora. In E2, benchmark scale reinforced textual bias. In E3, foundation models amplify textual dominance, sidelining oral and community knowledge.

6. **Soliciting data without reciprocal benefits**

Corpus tags: 16 (4% of all tags)

Compensation for data contributions has historically been absent or exploitative, especially for Indigenous and underserved groups [100, 195]. Labor and knowledge extraction without reciprocity creates longitudinal imbalances, such that contributors provide inputs but benefits flow elsewhere [88]. This pattern manifests across scales and contexts. For example, participatory design projects may create epistemic burdens on communities that remain unrecognized and uncompensated [163], or commercial platforms may treat user interactions as uncompensated "user-work" [32]. The absence of reciprocal arrangements transforms communities into resource suppliers rather than partners or beneficiaries (see 1 above), reinforcing extractive relationships where benefits concentrate in institutions and corporations while costs and risks are externalized to data contributors.

Era view. In E1, linguistic and anthropological projects often lacked reciprocity. In E2, crowd platforms offered transactional pay. In E3, foundation models scale contributions while benefits remain distant.

7. **Exploitative and invisible data labor**

Corpus tags: 14 (3% of all tags)

Data production for AI generates labor exploitation through both direct outsourcing and indirect harvesting of user activity. Sources document platform-mediated workers with limited wages and benefits processing raw data into labeled datasets and other training-ready formats [73]. Annotation and content moderation tasks are frequently outsourced to vulnerable populations in the Global Majority, where workers face repetitive and taxing environments with little protection or support when exposed to harmful content [132]. A parallel form of exploitation occurs through "unwitting" labor, where digital interactions generate training data without consent, communication of intent, or compensation [137]. Various systems extract free labor through mechanisms that disguise work as a necessary step to access, use, or improve online content: verification systems may double as annotation tasks, social media interactions may become sentiment analysis data, and user corrections may train autocomplete features [32**?**].

Era view. In E1, annotation was carried out by experts on small corpora. In E2, crowdwork scaled precarious conditions. In E3, unpaid and underpaid labor underpin foundation models, often obscured.

8. **Biased pre-processing and category erasure**

Corpus tags: 18 (4% of all tags)

Processing shortcuts embed erasures into dataset design. Biased categorization decisions begin during product conception and become operationalized through collection and enrichment practices that erase identities and contexts. Category formation practices often collapse distinctions like race/ethnicity, leading to the "lazy" erasure of identity categories and communities [190]. Annotation frameworks lacking community-defined criteria and annotator variance both precipitate downstream harms as they inject bias into collected data [62, 69, 85, 93]. This variance manifests in different ways depending on context, such as the omission of faith-based, religious, and spiritual values in postcolonial regions where they are central to social life [173]. Automated filtering and

cleaning pipelines compound these exclusions, for instance by treating disability identity as impermissible content [123] or by disproportionately removing data from non-Western contexts [81].

Era view. In E1, expert categories set narrow assumptions. In E2, large-scale annotation normalized shortcuts. In E3, automated pipelines amplify exclusions at scale.

9. **Keeping communities in the dark through opaque data practices**

Corpus tags: 15 (4% of all tags)

Opaque data practices deny communities access to their own data and information collected about them, which limits their ability to shape knowledge creation or contest outcomes [72, 95, 178]. Opacity manifests across multiple dimensions: technical complexity that obscures system functioning, surveillance-driven business models that hide how information is processed and monetized, and organizational practices that limit user knowledge of data sharing arrangements [112]. Legal restrictions further limit information availability, as do documentation gaps, leaving many AI datasets circulating without adequate provenance tracking or contextual information [116].

Era view. In E1, datasets circulated within institutional silos. In E2, benchmarks widened access but rarely tracked provenance. In E3, scraped corpora circulates with little documentation.

4.1.3 Deployment and Impact.

11. **Ethics dumping in less-regulated contexts**

Corpus tags: 9 (2% of all tags)

Ethics dumping describes the export of harmful or unethical research practices to underserved populations or to low- and middle-income countries where regulatory and ethical protections are weaker [135, 204]. Sources characterize this as transforming data into commodities with limited community return configuration. Fragmented regulatory frameworks permit foreign companies to harvest and repurpose citizen data without oversight or benefit-sharing, particularly in regions where legal protections remain underdeveloped [180]. Vulnerable countries become testing grounds for AI systems with heightened risk and little regard for autonomy, where underserved populations receive limited services in return for extensive data harvesting [64, 144], are part of trials in less-regulated contexts before deployment elsewhere, and have their data captured by foreign firms without oversight or benefit-sharing [208].

Cases documented in news reports (not included in corpus) include coercive biometric collection in Ethiopian refugee camps [205], Chinese surveillance technology deployment across Africa as testing grounds for authoritarian control [108], and Facebook's Cambridge Analytica data extraction affecting millions without informed consent [217].

Era view. In E1, ethics dumping remained largely unexamined. In E2, outsourcing to less-regulated regions became normalized. In E3, competitive model development intensifies these practices.

12. **Deploying AI systems trained without local, contextual data**

Corpus tags: 16 (4% of corpus)

Systems trained on datasets that exclude data from the communities they are served to or imposed upon can harm communities. When linguistic, cultural, demographic, or situational data from a community is absent,

deployed systems shift performance risks onto those very populations. The issue is not lack of data overall, but lack of contextually relevant data representing the diverse communities, which results in mismatches between system capabilities and user contexts [8]. The consequent harms range from quality-of-service failures to deeper representational exclusions, depending on the context and domain [189]. The deployment of healthcare AI tools lacking data from underserved populations in those same regions, for example, precipitates diagnostic failures and perpetuates health disparities [4, 18]. Language models trained primarily on North American and European data and deployed globally fail to serve local linguistic needs, and undermine community agency and cultural autonomy [193]. Systems developed without disability community input produce outputs that perpetuate subtle-yet-pervasive biases [69]. These deployment patterns prioritize market reach over contextual fit and shift performance risks onto underserved communities.

Era view. In E1, limited datasets constrained reach. In E2, global deployment expanded without proportional representation. In E3, foundation models magnify mismatches between training and deployment contexts.

13. **Western-centric research infrastructures**

Corpus tags: 7 (2% of corpus)

Research infrastructures shape AI data production by establishing which epistemologies and populations become legitimate sources of training data and evaluation criteria. Conferences, benchmarks, and publication venues concentrate in Western institutions, creating feedback loops where Euro-American categories become universal defaults while non-Western perspectives are systematically displaced [102]. Fairness research demonstrates this mismatch: 84% of studies rely exclusively on Western participants, yet resulting models reach global markets [188]. Industry partnerships compound exclusions through access barriers and non-disclosure agreements that prevent external scrutiny of how Western-centric assumptions shape data practices [181]. Academic structures embed extractive dynamics by treating knowledge as commodity while elevating Western perspectives and institutions over alternative epistemologies [101]. The infrastructure produces datasets that reflect institutional priorities concentrated in the Global North while marginalizing knowledge systems from the Global Majority.

Era view. In E1, expert curation reflected Western institutions. In E2, benchmarks and conferences reinforced Northern dominance. In E3, concentration of Western-centric viewpoints continues despite global deployment.

4.2 Less-extractive alternatives

Table 2 lists the eleven less-extractive pathways we identified across the pipeline. Counts represent coding tags rather than unique sources, since individual works may map to multiple mechanisms.

4.2.1 Problem Understanding and Formulation.

1. **Decentering Western ontologies**

Corpus tags: 27 (26 principles / 1 practice, 26:1)

The act of decentering Western ontologies redefines AI data production by shifting authority away from universalized categories toward situated worldviews. This principle positions diverse epistemologies as legitimate foundations for data practices without requiring conformity to Western categories. Indigenous and

Table 2. Eleven equitable data production pathways identified from corpus analysis across the AI development pipeline

Pipeline stage	Pipeline sub-stage	Less-extractive principles & practices
Problem understanding & formulation	Institutional prioritization & funding	1. Decentering Western ontologies
	Product conception & design	2. Taking a needs-based approach to developing AI 3. Early co-design and participatory initiatives
ML system design & development	Data selection, collection & annotation	4. Establishing consent and contextually appropriate compensation 5. Creating culturally inclusive datasets 6. Community engaged data production 7. Crowdsourcing data collection
	Model architecture selection & design	8. Building public visibility in dataset development 9. Developing equitable data licensing models 10. Developing federated data spaces
Pipeline-wide less-extractive pathways (principles & practices)		11. Participatory data ownership and governance

decolonial scholarship establishes that situated knowledge emerges from specific viewpoints and experiences rather than detached objectivity [192]. Sovereignty-based approaches prioritize reciprocity and accountability through "reverse tutelage" processes where technologists learn within community protocols before engaging with data [113, 135]. Relational ethics frameworks challenge individualistic assumptions by centering interdependence and collective responsibility [28, 131].

Examples include the JSwarm communication language that draws on Jingulu epistemologies for human-AI teaming systems [6], the Abundant Intelligences project that reimagines AI development through Indigenous knowledge systems [114], and decolonial big data psychology frameworks developed in Philippine contexts [136]. This principle applies across data types by positioning cultural context and community authority as foundational to collection decisions rather than technical afterthoughts. Scaling remains challenging as benchmark standardization continues to displace diverse epistemologies, though community-led initiatives demonstrate viable alternatives at regional levels [74, 178].

Era view. In E1, localized alternatives surfaced in small projects. In E2, benchmark standardization displaced them. In E3, community initiatives re-embed plural epistemologies, often at small scale.

2. **Taking a needs-based approach to developing AI**

Corpus tags: 7 (2 principles / 5 practices, 0.4:1)

When communities set priorities as the foundation for AI development, data production shifts from resource accumulation to community-driven innovation. This approach reverses conventional development by identifying what communities actually want before designing technical solutions. Communities often prioritize language preservation, cultural documentation, and local capacity building rather than conversational agents or mobile assistants that technologists typically assume [185]. Implementation requires sustained community engagement to understand local contexts and translate community goals into technical specifications. Needs assessments

and priority-setting exercises precede design decisions, aligning data collection and system development with community-defined objectives instead of external assumptions about utility.

Two examples illustrate this approach: writing assistants and translators co-designed with Brazilian Indigenous communities (Guarani Mbya and Nheengatu speakers) to address their expressed needs for language documentation, preservation, and vitalization, particularly for young people and content creators, with community data sovereignty maintained throughout [164]; and competitive multilingual models for low-resource African languages built effectively with small, curated localized datasets under 1 GB, addressing local speakers' needs for relevant, high-quality tools while mitigating ethical and environmental impacts of massive data extraction [148].

Era view. In E1, community needs were largely absent. In E2, pilots demonstrated feasibility but remained small. In E3, needs-based methods are increasingly framed as counterweights to global models.

3. **Early co-design and participatory initiatives**

Corpus tags: 40 (28 principles / 12 practices, 2.3:1)

When communities participate from the earliest stages of AI development, expert-driven design gives way to collaborative creation. Communities shape technical systems rather than adapt to predetermined functionality, and community knowledge and priorities are positioned as foundational inputs to problem formulation, system design, and data collection. Implementation operates through collaborative frameworks where community members function as co-designers rather than data subjects, establishing shared decision-making authority over technical specifications and development priorities. Projects require sustained engagement structures such as steering committees, community advisory boards, and iterative feedback mechanisms that maintain community influence throughout development cycles [152]. Participatory initiatives adapt co-design traditions to AI pipelines by making lived experience part of how datasets are scoped, labeled, and governed rather than treating community input as validation for predetermined technical decisions. Examples include the Masakhane organization's participatory research model for African language machine translation, where native speakers function as collaborators throughout dataset creation and model evaluation [142], participatory ethnographic and user experience research in African contexts with AI and speech technology that demonstrates the validity of community-centered approaches [57], and a feminicide counterdata collection initiative in Latin America that employs intersectional feminist principles to guide system design [198]. Community-led organizations such as Queer in AI further demonstrate how decentralized, volunteer-run initiatives can reshape AI development through participatory methods and intersectional analysis [168].

Era view. In E1, participatory methods were absent. In E2, pilot projects showed potential but limited reach. In E3, frameworks are more formalized, though tensions with commercial timelines persist.

4.2.2 ML System Design and Development .

4. **Establishing consent and contextually appropriate compensation**

Corpus tags: 11 (6 principles / 5 practices, 1.2:1)

Dynamic consent frameworks give communities ongoing control over data use and fair compensation aligned with their priorities. Agreements require renegotiation during and after collection to address changing uses and circumstances [176]. Compensation aligns with community priorities rather than standardized rates and

may include monetary and non-monetary benefits such as attribution, capacity building, and community benefit sharing. Incentive structures can reward accurate contributions while protecting contributor autonomy. Examples include partnerships with disability communities where participants accepted lower monetary compensation when data improved community life [156], Indigenous research frameworks that prioritize respect through attribution and capacity building [103], and incentive-compatible mechanisms for fitness data that compensate quality without surveillance [209]. Proposed practices include embodied consent frameworks adapted for vulnerable groups in AI contexts that emphasize freely given, reversible, informed, enthusiastic, and specific agreements [196], and levy systems for generative AI that would compensate copyright owners through streamlined opt-out mechanisms [157].

Era view. In E1, contributors rarely received compensation. In E2, standardized payments prioritized efficiency. In E3, lifecycle-based and context-sensitive models gain traction.

5. **Creating culturally inclusive datasets**

Corpus tags: 45 (3 principles / 42 practices, 0.07:1)

When datasets reflect community priorities and cultural contexts, representation moves beyond numerical coverage to preserve meaning and agency. Cultural frameworks become foundational to design, resulting in data production that safeguards context and guides appropriate use. Implementations range from intentionally small collections that prioritize cultural depth to regional collaborations that scale while maintaining community control. Examples of small collections include Arabic proverb datasets capturing figurative language across dialects [124], and AAVE corpora centering Black speech patterns [79, 222]. Regional initiatives demonstrate community-driven approaches through, for instance, African language model development [9, 210] and data collection platforms in rural India that ensure fair wages and worker ownership [94]. Cross-cultural evaluation frameworks demonstrate how geographic and demographic perspectives shape dataset quality, for example multicultural alignment datasets that capture diverse human preferences [96] and community-in-the-loop benchmarks for underrepresented groups [66]. The practice transforms AI data production from neutral accumulation toward cultural presence and accountability where communities control how their knowledge is represented and used.

Era view. In E1, expert corpora reflected narrow perspectives. In E2, benchmark coverage expanded but sidelined meaning. In E3, inclusive corpora range from small intentional datasets to regional collaborations.

6. **Community engaged data production**

Corpus tags: 39 (11 principles / 28 practices, 0.4:1)

Community-engaged data production positions communities as custodians and creators, not subjects. Communities retain agency over what gets recorded, how materials are stored, and how benefits circulate through negotiated agreements that adapt to local decision-making structures instead of imposing external frameworks [24, 44, 120]. Community members function as technical collaborators who bridge linguistic knowledge and implementation decisions. Collection methods align with community priorities rather than external assumptions [26, 221]. Technical workflows adapt to local knowledge validation practices, from elder councils to distributed networks.

Language data production (i.e., text and speech) occurs through trust-based partnerships where communities define evaluation criteria and success metrics according to their own goals, such as language revitalization or

cultural preservation, and reject standardized benchmarks (see 2). Examples include Māori communities developing culturally appropriate benchmarks through Te Hiku Media [60], Lakota/Cheyenne/Blackfeet communities generating community-defined datasets while maintaining data sovereignty [127], and Ugandan communities creating language corpora through local NLP teams [12]. The Speech Accessibility Project engages disability communities as partners in data production: organizations collect speech samples from paid volunteers while communities review consent processes and identify needs [2].

Visual data production follows similar principles, with communities redefining images as cultural artifacts rather than neutral data points, from Hawai'ian self-representation archives [175] to food datasets stewarded through ambassador networks and expert review [76]. Large-scale initiatives demonstrate how community engagement can scale across regions while preserving local autonomy [3, 149]. The practice transforms data production from resource extraction to collaborative creation where technical capacity building and community empowerment take precedence over dataset accumulation.

Era view. In E1, experts curated corpora with little community input. In E2, participatory pilots introduced shared methods. In E3, community-engaged projects establish datasets as cultural resources, though often resource-intensive.

7. **Crowdsourcing data collection**

Corpus tags: 12 (0 principles / 12 practices, ∞:1 practice-heavy)

Crowdsourcing data collection mobilizes distributed contributions to build datasets. Outcomes vary dramatically based on how platforms structure ownership, compensation, and governance. Less-extractive approaches operate through open infrastructures that enable shared corpus development and simultaneously maintain contributor recognition. Decentralized models use blockchain-based verification to align data provenance with contributor sovereignty in marketplaces where crowds retain ownership while allowing computational access to their data [50, 119].

Examples span multiple models of community engagement and benefit distribution. Mozilla Common Voice demonstrates volunteer-driven preservation of endangered languages that likewise challenges the commercial capture of speech data [212]. Microsoft Research's partnerships with low-income Indian communities show how economic frameworks can provide fair compensation for crowdsourced data toward the aim of producing high-quality culturally inclusive datasets [7]. The World Wide Dishes project illustrates heritage-based networks where volunteers shared knowledge about their own cultural cuisines through social media mobilization, with community ambassadors facilitating participation and distinguished contributors becoming co-authors [126]. Ubuntu-AI represents ownership-based systems that enable African artists to retain control and receive profit-sharing from AI training rather than facing displacement [141].

Era view. In E1, crowdsourcing was absent. In E2, platforms expanded participation but with limited governance. In E3, decentralized models connect crowdsourcing to sovereignty, though scale is uncertain.

8. **Building public visibility in dataset development**

Corpus tags: 21 (14 principles / 7 practices, 2:1)

Documentation practices have become a recognized accountability mechanism in AI development, creating standardized formats that disclose dataset properties. Early experiments with "nutritional labels" [80], representation summaries [23], and ranking algorithm disclosures [219] laid the groundwork. Model cards [134] and

datasheets [72, 83] consolidated these approaches into formats that are now widely adopted in industry settings. Yet standardization primarily addresses the needs of technical teams and regulators, offering implementation details for specialized audiences rather than accountability information accessible to affected communities [46]. Standard formats also miss the day-to-day decision-making that shapes datasets [218].

Alternative approaches address accountability gaps in different ways. Methodological reforms adapt archival practices for systematic transparency frameworks [90] and replace static documentation with iterative quality measurement [85]. Audience-focused reforms integrate community empowerment considerations into technical specifications [155] and test broader accessibility formats through municipal pilots like Helsinki [146](not in corpus). Enforcement mechanisms show how public disclosure can turn into corporate pressure: Gender Shades, a research project that publicized racial and gender bias in AI facial recognition systems, led measurable improvements when this biased performance was made visible [170].

Era view. In E1, documentation was minimal and expert-facing. In E2, structured formats emerged with benchmarks. In E3, transparency is institutionalized, though accessibility and impact are unevenly distributed.

9. **Developing equitable data licensing models**

Corpus tags: 3 (1 principle / 2 practices, 0.5:1)

When communities design legal terms that make their authority a precondition for dataset access, they embed sovereignty and worker protection into governance frameworks. Licenses restrict harmful applications, shield annotators from exploitation, return benefits to contributors, and replace open-source defaults or proprietary terms with community-controlled conditions [151].

For example, the Kaitiakitanga License requires alignment with Māori tikanga (customs and protocols) and ensures Māori communities benefit from their own data rather than merely providing inputs for external systems [203]. The Nwulite Obodo Open Data License balances openness with community benefit for African language datasets. The Esethu Framework establishes sovereignty provisions for isiXhosa speech data as well as protective provisions for annotators and community benefit-sharing [169].

Era view. In E1, academic norms favored open use. In E2, open-source defaults prioritized access over sovereignty. In E3, community-designed licenses embed sovereignty, though uptake is limited.

10. **Developing federated data spaces**

Corpus tags: 4 (0 principles / 4 practices, ∞:1 practice-heavy)

Federated data spaces redistribute data production by allowing collaboration without centralizing control. The space model entails decentralized infrastructures with technical and governance frameworks that support trust between participants, sovereignty for data providers, and transparency in handling [40, 107]. Implementation requires sovereignty guarantees, clear governance rules, robust security protocols, and interoperability standards that allow organizations to share computational access yet retain data ownership [84]. The EU's Data Spaces Support Centre explores how federated models can provide secure, governed infrastructure for high-quality AI training data and at the same time foster regulatory compliance and new business opportunities [70]. Despite technical complexity, implementation costs, and regulatory uncertainty, these largely proposed models offer potential pathways for collaborative AI development that preserve organizational and community autonomy over data resources.

Era view. In E1, sharing required full transfer of control. In E2, platforms centralized authority. In E3, federated models propose redistribution, though mostly at pilot stage.

4.2.3 *Pipeline-wide less-extractive principles and practices.*

11. **Participatory data ownership & governance**

Corpus tags: 44 (37 principles / 7 practices, 5.3:1)

Governance frameworks reshape data production by embedding sovereignty, justice, and accountability into decisions about collection, access, and use. Broader currents in data justice foreground consent as a revocable, user-controlled right [11] and redirect value through generative approaches that emphasize reciprocity [63, 202]. Indigenous data sovereignty centers governance in cultural protocols and collective responsibility [37, 106, 113]. Communities bring distinct priorities, histories, and epistemologies. They are not monolithic and resist one-size-fits-all oversight and data management techniques.

The principles of sovereignty and accountability take concrete form when communities themselves organize to set terms for how data is governed. The limited integration of Indigenous perspectives into mainstream policymaking has led communities worldwide to self-organize and advocate for their unique concerns, especially in the current age of AI. One key example is the Māori community in New Zealand, where the Māori Data Sovereignty Network (MDSN) has emerged as a crucial voice [117]. This network comprises dedicated professionals—including academics, researchers, data scientists, and community leaders—who are committed to safeguarding the interests of the Māori people, particularly in the digital realm. They advocate for Māori data self-determination, ensuring ethical governance and culturally appropriate practices that align with traditional Māori values and principles [105, 138].

This proactive approach is not unique to the Māori. Across Africa, the Masakhane policy group demonstrates a similar commitment to self-advocacy. This dynamic group has launched significant policy initiatives aimed at advancing the recognition and use of African languages in various domains, particularly within the context of AI. Masakhane's advocacy extends to championing appropriate governmental policies that foster the development and deployment of AI technologies in a manner that respects linguistic diversity, promotes local innovation, and ensures equitable access and benefits for African communities [128]. Both the MDSN and Masakhane exemplify the critical role of Indigenous and community-led organizations in filling the void left by insufficient governmental inclusion, thereby shaping policies that are culturally sensitive, equitable, and truly representative of their peoples' needs and aspirations.

Era view. In E1, governance was centralized in institutions. In E2, protections focused on narrow consent. In E3, community-led initiatives embed sovereignty into policy and infrastructure.

5 Discussion

5.1 Upstream intervention imperative

Table 3 reports orientation counts in unique sources across pipeline stages. Extractive patterns concentrate in ML System Design and Development, where just over half (73 of 142 cases, 51%) are located. This stage also holds the

Table 3. Distribution of orientations across AI pipeline stages, reported in *unique sources*. Extractive patterns cluster in ML System Design & Development (73 of 142 unique sources, 51%), while less-extractive principles and practices appear more evenly distributed. Counts indicate the number of sources primarily coded to each stage.

Pipeline Stage	Extractive Patterns	Less-extractive Principles	Less-extractive Practices	Total # of Unique Sources	% Extractive Orientation	% Less-extractive Orientation*	% of Total Corpus
Problem Understanding & Selection	36	44	10	90	40%	60%	26%
ML System Design & Development	73	34	76	183	40%	60%	51%
Deployment & Impact	14	5	3	22	64%	36%	6%
Cross-pipeline	19	32	4	55	35%	65%	16%
Total	142	115	93	**350**	-	-	100%

* % Less-extractive Orientation = (principles + practices) / Total for the row.
% Extractive Orientation = Extractive patterns / Total for the row.
Note: Percentages are row-wise. For example, 73 of 142 extractive sources (51%) occur in ML System Design & Development.

highest number of less-extractive practices (76 unique sources). Technical development thus operates as both the main site of extraction and the most active space for alternatives in AI data production.

Interventions differ in difficulty across the pipeline. Early-stage change demands shifts in funding priorities and problem formulation, which are controlled by institutions with concentrated power. Later-stage responses often reduce to narrow technical adjustments that do not alter those upstream arrangements. Once architectures and evaluation metrics are set, community participation is reduced to annotation or consultation, with little ability to shape fundamental choices.

These dynamics place the A/D/C triangle under structural strain. AI systems (A) and data production practices (D) drive decisions, with community impacts (C) addressed only after design trajectories are fixed. The evidence points to upstream decision-making as the most consequential intervention point for the direction of AI data production as a whole. Without shifts at this stage, the triangle stays unbalanced, with C treated as an afterthought rather than a constitutive vertex of development. The concentration of extractive patterns in technical phases suggests that technical solutions alone cannot address systemic problems rooted in institutional priorities.

5.2 Action gap and evidence variation

Table 4 reports principles and practices as coding tags. Ratios show a consistent action gap: for every practice, 2.7 sources advance critique or principles (257 vs. 93). Principles also outnumber practices overall at 1.2:1. This unevenness stretches the A/D/C triangle: critique and normative work concentrate around AI systems (A) and data practices (D), but far fewer practices connect directly to community impacts (C).

Pathways differ in emphasis and implementation difficulty. Decentering Western ontologies stands at 26:1, indicating a strong tilt toward principle-setting. Participatory governance is also principle-heavy at 5.3:1. In contrast,

Table 4. Principles-to-practices ratios for each less-extractive pathway, reported in *coding tags*. Lower ratios reflect stronger documentation of practices; higher ratios indicate pathways that remain largely theoretical. Ratios also highlight intervention scale: technical methods show easier implementation, whereas governance frameworks require broader institutional change.

Less-extractive Pathway	Principles	Practices	Ratio (P:Pr)	Implementation Priority
Decentering Western ontologies	26	1	26:1	High need
Participatory data ownership and governance	37	7	5.3:1	High need
Building public visibility in dataset development	14	7	2:1	Moderate need
Early co-design and participatory initiatives	28	12	2.3:1	Moderate need
Establishing consent and contextually appropriate compensation	6	5	1.2:1	Balanced
Developing equitable data licensing models	1	2	0.5:1	Balanced
Community engaged data production	11	28	0.4:1	Well-implemented
Taking a needs-based approach to developing AI	2	5	0.4: 1	Well-implemented
Creating culturally inclusive datasets	3	42	0.07:1	Well-implemented
Crowdsourcing data collection	0	12	∞:1	Practice-heavy
Developing federated data spaces	0	4	∞:1	Practice-heavy

crowdsourcing and culturally inclusive dataset creation lean heavily toward practices, suggesting that technical interventions aligned with existing infrastructures encounter fewer adoption barriers. Yet interventions differ in scale and impact: crowdsourcing data collection is comparatively simple to implement, whereas governance frameworks or sovereignty-based licensing demand broader institutional change.

Geographic concentration adds another layer. Successful alternatives cluster in specific ecosystems (e.g., African NLP networks such as Masakhane, Māori data sovereignty frameworks, the Indigenous-led, Indigenous-majority international, interdisciplinary research program Abundant Intelligence) rather than diffusing across institutional contexts. This clustering suggests that less-extractive pathways emerge from specific cultural and political conditions rather than universal principles that can be deployed everywhere.

The action gap reflects structural alignment of incentives with extraction. Less-extractive pathways require parallel infrastructures, alternative evaluation metrics, and economic models that redistribute rather than concentrate benefits. Without these conditions, the triangle tilts toward analysis and principle-setting, with community impacts pulled to the margins instead of anchored as a constitutive vertex of development.

5.3 Ecosystem emergence

Landscape analyses compile individual sources and then step back to identify patterns across the corpus. Here, one striking pattern is emergent ecosystems: diverse initiatives by different actors and groups all driving toward the same vision of equitable AI data production. Authors, organizations, and initiatives offer unique contributions, but their coordination and totality points to systemic alternatives that start with community needs and maintain community control. Like healthy ecosystems in nature, these interconnected initiatives create their own conditions for thriving rather than adapt to external constraints. They develop their own evaluation criteria, publication venues, funding mechanisms, and governance protocols. They establish parallel infrastructures that operate according to different

principles: sovereignty rather than extraction, reciprocity rather than accumulation, cultural preservation rather than homogenization and standardization.

The African context illustrates this dynamic clearly. African NLP networks such as Masakhane, born out of the Deep Learning Indaba, drive agenda-setting for continent-wide initiatives [49]. Sunbird AI connects Ugandan social challenges to practical AI applications, Ghana NLP focuses on West African language inclusion, and Abena AI scales speech technologies across linguistic communities. InkubaLM by Lelapa AI exemplifies purpose-oriented data production rooted in local languages and applications, where model development connects directly to social challenges and regional talent [216]. The African Next Voices project represents a massively scaled approach, the largest community-driven language data initiative to date, that engages directly with diverse communities to produce data for AI rather than scrape existing content [220].

The corpus registers the continent's ecosystem as it shifts from critique to action. Fifteen African sources diagnose extractive patterns (ethics dumping, digital extractivism, governance gaps), but twenty-six sources advance solutions through practices and principles. The practice sources concentrate on culturally inclusive datasets (12 sources), community-engaged data production (8 sources), and participatory approaches, addressing technical infrastructure (language models, datasets), organizational infrastructure (community-led organizations across Uganda, Kenya, Ghana, South Africa), legal infrastructure (community-centric licensing), and methodological infrastructure (participatory research frameworks). Sources reflect a 26:15 ratio of forward-looking to diagnostic work, and the coordination of universities, NGOs, industry actors, and independent organizations across the continent.

The pattern extends beyond Africa. Māori data sovereignty frameworks in Aotearoa New Zealand similarly demonstrate coordinated ecosystem approaches: the Māori Data Sovereignty Network develops governance protocols, Te Hiku Media creates community-led, culturally appropriate datasets and benchmarks, and the Kaitiakitanga License embeds community authority into legal frameworks. Like the African examples, these initiatives operate across technical, organizational, and legal infrastructures while maintaining cultural specificity.

The pattern is not unique to regions, either. Global community-led organizations like Abundant Intelligences [114], Black in AI [20], LatinX in AI [122], Queer in AI [166], and the Speech Accessibility Project [2] operate across continents and institutional boundaries. Some, like the Speech Accessibility Project, directly drive data production data for AI; others convene collaborative venues, establish alternative publication pathways, and center community priorities over technical solutions—all efforts similarly integral to data production for AI.

Ecosystemic movements of all shapes and scales redefine AI data production end-to-end by creating outcomes that exceed the sum of individual parts and invert conventional development logic. Instead of communities adapting to predetermined technical systems, ecosystems adapt technical development to community needs and governance structures. Community knowledges and needs replace institutional priorities across the entire pipeline, from agenda-setting to benefit distribution.

6 Implications for HCI research and practice

Four specific implications follow for HCI researchers and practitioners:

(1) **Expand the unit of analysis**. Data production encompasses institutional funding decisions, problem formulation, and agenda-setting processes that determine whose knowledge becomes computational input. HCI methods should account for these upstream processes rather than treating them as external constraints. Research should investigate how power circulates through data production networks, examining who controls technical development and how benefits are distributed.
(2) **Support ecosystem coordination rather than isolated projects**. Rather than defaulting to standardized approaches that may flatten nuance or displace local epistemologies, HCI methods should highlight and empower collaboration across contexts and disciplines.
(3) **Address structural implementation barriers**. The 2.7:1 action gap and geographic clustering of practices indicate that normative frameworks are insufficient for change. Research should investigate how existing technical infrastructures, economic arrangements, and institutional incentives reward extraction and shape which interventions are possible and which alternatives can be sustained.
(4) **Recognize ecosystem-level coordination as a research priority**. Community-led ecosystems create systemic alternatives that exceed individual project impacts. HCI research should identify, analyze, and support these coordinated efforts rather than treating them as isolated cases. This requires methods that can capture coordination across organizations, regions, and technical domains.

7 Conclusions

Our landscape analysis of 350 sources on AI data production—the processes that transform human knowledge into computational resources—identified twelve extractive patterns and eleven less-extractive pathways across development pipeline stages and historical eras. Sources engaged artificial intelligence systems (A), data production practices (D), and community impacts (C) in varying combinations. Extractive patterns concentrate in technical development phases (51% in ML System Design & Development). This same stage hosts nearly equal numbers of less-extractive practices (76 practices vs. 73 extractive patterns)—a critical finding given the action gap between principles and implementation documented across the literature. Across the pipeline, community-led ecosystems create coordinated alternatives that generate systemic change beyond individual projects.

Three findings reshape understanding of AI data production. First, upstream intervention proves critical: technical decisions about AI systems and data practices occur without meaningful community input, leaving communities to deal with consequences rather than shape development. Second, a substantial action gap persists: for every documented practice, nearly three sources advance critique or principles, indicating structural barriers beyond awareness or frameworks. Third, ecosystem emergence offers pathways forward: coordinated networks in Africa, Māori data sovereignty initiatives, and global community organizations demonstrate how alternatives can maintain community control across technical, organizational, and legal infrastructures.

Moving from diagnosis to action requires addressing fundamental questions: How do community-controlled datasets, sovereignty-based licensing models, and participatory governance frameworks interact to create ecosystem-level alternatives? What infrastructures support decentralized, community-led AI data production at scale? HCI should examine its own role in either reproducing or disrupting extractive patterns through emphasis on innovation, efficiency, and scalability that may treat community knowledge as raw material.

Our systematic mapping shows that interventions must address upstream decision-making, implementation barriers, and ecosystem coordination simultaneously. AI data production will expand globally. Whether it operates through extraction or reciprocity depends on deliberate choices about whose knowledge counts, who controls technical development, and how benefits circulate. Both trajectories are possible. The question is which communities, institutions, and technical systems will shape that direction.

References

[1] [n. d.]. Data Workers Inquiry. https://data-workers.org/

[2] [n. d.]. Speech Accessibility Project. https://speechaccessibilityproject.beckman.illinois.edu

[3] [n. d.]. Vaani. https://vaani.iisc.ac.in/

[4] 2023. *Responsible AI in Africa: Challenges and Opportunities*. Springer International Publishing, Cham. doi:10.1007/978-3-031-08215-3

[5] 2024. Data Provenance Initiative. https://www.dataprovenance.org/

[6] Hussein A. Abbass, Eleni Petraki, and Robert Hunjet. 2022. JSwarm: A Jingulu-Inspired Human-AI-Teaming Language for Context-Aware Swarm Guidance. *Frontiers in Physics* 10 (2022). doi:10.3389/fphy.2022.944064

[7] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2819–2826. https://aclanthology.org/2020.lrec-1.343/

[8] Rachel Adams, Ayantola Alayande, Zameer Brey, Brantley Browning, Michael Gastrow, Jerry John Kponyo, Dona Mathew, Moremi Nkosi, Henry Nunoo-Mensah, Diana Nyakundi, Victor Odumuyiwa, Olubunmi Okunowo, Philipp Olbrich, Nawal Omar, Kemi Omotubora, Paul Plantinga, Gabriella Razzano, Zara Schroeder, Andrew Selasi Agbemenu, Araba Sey, Kristophina Shilongo, Shreya Shirude, Matthew Smith, Eric Tutu Tchao, and Davy K. Uwizera. 2023. A new research agenda for African generative AI. *Nature Human Behaviour* 7, 11 (Nov. 2023), 1839–1841. doi:10.1038/s41562-023-01735-1

[9] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics* 9 (Oct. 2021), 1116–1131. doi:10.1162/tacl_a_00416

[10] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3613904.3642703

[11] Leah Ajmani, Logan Stapleton, Mo Houtti, and Stevie Chancellor. 2024. Data Agency Theory: A Precise Theory of Justice for AI Applications. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 631–641. doi:10.1145/3630106.3658930

[12] Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine Translation for African Languages: Community Creation of Datasets and Models in Uganda. *AfricaNLP workshop at ICLR2022* (2022). https://openreview.net/pdf?id=BK-z5qzEU-9

[13] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12404–12422. doi:10.18653/v1/2024.acl-long.671

[14] Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar AL-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Anfel Rouabhia, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 32871–32894. doi:10.18653/v1/2025.acl-long.1579

[15] Marianna Anagnostou, Olga Karvounidou, Chrysovalantou Katritzidaki, Christina Kechagia, Kyriaki Melidou, Eleni Mpeza, Ioannis Konstantinidis, Eleni Kapantai, Christos Berberidis, Ioannis Magnisalis, and Vassilios Peristeras. 2022. Characteristics and challenges in the industries towards responsible AI: a systematic literature review. *Ethics and Information Technology* 24, 3 (2022), 37. doi:10.1007/s10676-022-09634-1

[16] Hilary Arksey and Lisa O'Malley. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 8, 1 (Feb. 2005), 19–32. doi:10.1080/1364557032000119616

[17] A. Arora, M. Barrett, E. Lee, E. Oborn, and K. Prince. 2023. Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization* 33, 3 (2023), 100478. doi:10.1016/j.infoandorg.2023.100478

[18] Mercy Asiedu, Awa Dieng, Iskandar Haykel, Negar Rostamzadeh, Stephen Pfohl, Chirag Nagpal, Maria Nagawa, Abigail Oppong, Sanmi Koyejo, and Katherine Heller. 2024. The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa. arXiv:2403.03357 (March 2024). doi:10.48550/arXiv.2403.03357 arXiv:2403.03357 [cs].

[19] Gelan Ayana, Kokeb Dese, Hundessa Daba Nemomssa, Bontu Habtamu, Bruce Mellado, Kingsley Badu, Edmund Yamba, Sylvain Landry Faye, Moise Ondua, Dickson Nsagha, Denis Nkweteyim, and Jude Dzevela Kong. 2024. Decolonizing global AI governance: assessment of the state of decolonized AI governance in Sub-Saharan Africa. *Royal Society Open Science* 11, 8 (Aug. 2024), 231994. doi:10.1098/rsos.231994

[20] BAI. 2017. Black in AI. https://www.blackinai.org/

[21] Teanna Barrett, Chinasa T. Okolo, B. Biira, Eman Sherif, Amy Zhang, and Leilani Battle. 2025. African Data Ethics: A Discursive Framework for Black Decolonial AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 334–349. doi:10.1145/3715275.3732023

[22] Marie Battiste. 2005. Indigenous Knowledge: Foundations for First Nations. *Worm Indigenous Nations Higher Education Consortium Journal* (Jan. 2005). https://www.researchgate.net/publication/241822370_Indigenous_Knowledge_Foundations_for_First_Nations

[23] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), 587–604. doi:10.1162/tacl_a_00041

[24] Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 3504–3519. doi:10.18653/v1/2020.coling-main.313

[25] Steven Bird. 2024. Must NLP be Extractive? https://drive.google.com/file/d/1hvF7_WQrou6CWZydhymYFTYHnd3ZIljV/view?usp=embed_facebook

[26] Steven Bird and Dean Yibarbuk. 2024. Centering the Speech Community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 826–839. https://aclanthology.org/2024.eacl-long.50

[27] Abeba Birhane. 2020. Algorithmic Colonization of Africa. *SCRIPTed* 17, 2 (Aug. 2020), 389–409. doi:10.2966/scrip.170220.389

[28] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (Feb. 2021), 100205. doi:10.1016/j.patter.2021.100205

[29] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. doi:10.1145/3551624.3555290

[30] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. (2021). doi:10.48550/ARXIV.2110.01963

[31] Sibusiso Biyela, Amr Rageh, and Shakoor Rather. 2025. As AI giants duel, the Global South builds its own brainpower. *Natre [immersive feature]* (2025). https://www.nature.com/immersive/d44151-025-00085-3/index.html

[32] Briony Blackmore, Michelle Thorp, Andrew Tzer-Yeu Chen, Fabio Morreale, Brent Burmester, Elham Bahmanteymouri, and Matt Bartlett. 2023. Hidden humans: exploring perceptions of user-work and training artificial intelligence in Aotearoa New Zealand. *Kōtuitui: New Zealand Journal of Social Sciences Online* 18, 4 (Oct. 2023), 443–456. doi:10.1080/1177083X.2023.2212736

[33] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. doi:10.18653/v1/2020.acl-main.485

[34] Paul T. Brown, Daniel Wilson, Kiri West, Kirita-Rose Escott, Kiya Basabas, Ben Ritchie, Danielle Lucas, Ivy Taia, Natalie Kusabs, and Te Taka Keegan. 2024. Māori Algorithmic Sovereignty: Idea, Principles, and Use. *Data Science Journal* 23, 1 (April 2024). doi:10.5334/dsj-2024-015

[35] Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge. Google-Books-ID: kuztAAAAMAAJ.

[36] Deborah Cameron, Elizabeth Frazer, Penelope Harvey, Ben Rampton, and Kay Richardson. 1993. Ethics, advocacy and empowerment: Issues of method in researching language. *Language Communication* 13, 2 (April 1993), 81–94. doi:10.1016/0271-5309(93)90001-4

[37] Stephanie Russo Carroll, Pyrou Chung, Robyn K. Rowe, Susanna Siri, and Walter. 2024. Indigenous Data Sovereignty and the State of Open Data. https://www.d4d.net/news/indigenous-data-sovereignty-and-the-state-of-open-data

[38] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7 (April 2024), 166:1–166:38. doi:10.1145/3616865

[39] Ishita Chordia, Leya Breanna Baltaxe-Admony, Ashley Boone, Alyssa Sheehan, Lynn Dombrowski, Christopher A Le Dantec, Kathryn E. Ringland, and Angela D. R. Smith. 2024. Social Justice in HCI: A Systematic Literature Review. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–33. doi:10.1145/3613904.3642704

[40] NL AI Coalition. 2023. *Towards a federation of AI data spaces: NL AIC reference guide to federated and interoperable AI data spaces.* https://coe-dsc.nl/nl-aic-publishes-guide-towards-a-federation-of-ai-data-spaces/

[41] Cathy J. Cohen. 1997. Punks, Bulldaggers, and Welfare Queens: The Radical Potential of Queer Politics? *GLQ: A Journal of Lesbian and Gay Studies* 3, 4 (1997), 437–465.

[42] Combahee River Collective. 1977. (1977) The Combahee River Collective Statement •. https://www.blackpast.org/african-american-history/combahee-river-collective-statement-1977/

[43] Patricia Hill Collins. 2002. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment* (2 ed.). Routledge, New York. doi:10.4324/9780203900055

[44] Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. "It's how you do things that matters": Attending to Process to Better Serve Indigenous Communities with Language Technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 204–211. https://aclanthology.org/2024.eacl-short.19

[45] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–18. doi:10.1145/3491102.3517716

[46] Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3617694.3623228

[47] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature* 630, 8018 (2024), 841–846. doi:10.1038/s41586-024-07335-x

[48] Kate Crawford. 2021. Atlas of AI. https://yalebooks.yale.edu/9780300264630/atlas-of-ai

[49] Georgina Curto, Frank Dignum, Girmaw Abebe Tadesse, Avishkar Bhoopchand, Sibusisiwe Makhanya, Vukosi Marivate, and Emma Ruttkamp-Bloem. 2025. Africa leading the global effort for AI that works for all. *Nature Africa* (May 2025). doi:10.1038/d44148-025-00141-1

[50] Venkata Satya Sai Ajay Daliparthi, Nurul Momen, Kurt Tutschku, and Miguel De Prado. 2023. ViSDM 1.0: Vision Sovereignty Data Marketplace a Decentralized Platform for Crowdsourcing Data Collection and Trading. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good (GoodIT '23)*. Association for Computing Machinery, New York, NY, USA, 374–383. doi:10.1145/3582515.3609556

[51] Íñigo de Troya, Jacqueline Kernahan, Neelke Doorn, Virginia Dignum, and Roel Dobbe. 2025. Misabstraction in Sociotechnical Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1829–1842. doi:10.1145/3715275.3732122

[52] Jeffrey Dean. 2019. The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design. arXiv:1911.05289 (Nov. 2019). doi:10.48550/arXiv.1911.05289 arXiv:1911.05289 [cs].

[53] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3617694.3623261

[54] Vine Deloria and Clifford M. Lytle. 1998. *The Nations Within: The Past and Future of American Indian Sovereignty*. University of Texas Press. Google-Books-ID: FLgEf5kGLWQC.

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, 248–255. doi:10.1109/CVPR.2009.5206848

[56] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data Society* 8, 2 (2021), 205395172110359. doi:10.1177/20539517211035955

[57] Lindsey DeWitt Prat, Olivia Nercy Ndlovu Lucas, Christopher Golias, and Mia Lewis. 2024. Decolonizing LLMs: An Ethnographic Framework for AI in African Contexts. *EPIC Proceedings* (2024), 45–84. https://doi.org/10.1111/epic.12196

[58] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. arXiv:2104.08758 (2021). doi:10.48550/arXiv.2104.08758 arXiv:2104.08758 [cs].

[59] Mindy Duffourc, Sara Gerke, and Konrad Kollnig. 2024. Privacy of Personal Data in the Generative AI Data Lifecycle. 4899219 (2024). doi:10.2139/ssrn.4899219

[60] Suzanne Duncan, Gianna Leoni, Lee Steven, Keoni Mahelona, and Peter-Lucas Jones. 2024. Fit for our purpose, not yours: Benchmark for a low-resource, Indigenous language. https://openreview.net/forum?id=w5jfyvsRq3#discussion

[61] Francesco Durante, Markus Kröger, and Will LaFleur. 2021. *Extraction and Extractivisms*. Routledge, 17–30. doi:10.4324/9781003127611-3

[62] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351. doi:10.1145/3531146.3534647

[63] Ron Eglash, Kwame P Robinson, Audrey Bennett, Lionel Robert, and Mathew Garvin. 2024. Computational reparations as generative justice: Decolonial transitions to unalienated circular value flow. *Big Data Society* 11, 1 (March 2024), 20539517231221732. doi:10.1177/20539517231221732

[64] Brian Ekdale and Melissa Tully. 2019. African Elections as a Testing Ground: Comparing Coverage of Cambridge Analytica in Nigerian and Kenyan Newspapers. *African Journalism Studies* 40, 4 (Oct. 2019), 27–43. doi:10.1080/23743670.2019.1679208

[65] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 38–48. doi:10.1145/3600211.3604661

[66] Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. arXiv:2306.15087 (2023). doi:10.48550/arXiv.2306.15087 arXiv:2306.15087 [cs].

[67] Raymond Fok, Alexa Siu, and Daniel S. Weld. 2025. Toward Living Narrative Reviews: An Empirical Study of the Processes and Challenges in Updating Survey Articles in Computing Research. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3706598.3714047

[68] Miranda Fricker. 2007. *Epistemic injustice: power and the ethics of knowing*. Oxford university press, Oxford.

[69] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 205–216. doi:10.1145/3593013.3593989

[70] Ana García, Savvas Rogotis, Eimear Farrell, Tobias Guggenberger, Arash Hajikhani, Atte Kinnula, Marko Komssi, and Tuomo Tuikka. 2024. Generative AI and Data Spaces: White Paper. (2024).

[71] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (Feb. 2019), 101–121. doi:10.1016/j.infsof.2018.09.006

[72] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. doi:10.1145/3458723

[73] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

[74] Lara Groves, Aidan Peppin, Andrew Strait, and Jenny Brennan. 2023. Going public: the role of public participation approaches in commercial AI labs. arXiv:2306.09871 (2023). doi:10.48550/arXiv.2306.09871 arXiv:2306.09871 [cs].

[75] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (March 2009), 8–12. doi:10.1109/MIS.2009.36

[76] Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Sefala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, and Tejumade Afonja. 2025. The Human Labour of Data Work: Capturing Cultural Diversity through World Wide Dishes. arXiv:2502.05961 (Feb. 2025). doi:10.48550/arXiv.2502.05961 arXiv:2502.05961 [cs].

[77] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. doi:10.2307/3178066

[78] Paula Helm, Benjamin Lipp, and Roser Pujadas. 2024. Generating reality and silencing debate: Synthetic data as discursive device. *Big Data Society* 11, 2 (2024), 20539517241249447. doi:10.1177/20539517241249447

[79] Jazmia Henry. 2025. *aave_corpora*. https://github.com/jazmiahenry/aave_corpora Jupyter Notebook, MIT License.

[80] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 (May 2018). doi:10.48550/arXiv.1805.03677 arXiv:1805.03677 [cs].

[81] Sun-ha Hong. 2023. Prediction as extraction of discretion. *Big Data Society* 10, 1 (Jan. 2023), 20539517231171053. doi:10.1177/20539517231171053

[82] Ben Hutchinson. 2024. Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing. arXiv:2404.14740 (2024). doi:10.48550/arXiv.2404.14740 arXiv:2404.14740 [cs].

[83] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 560–575. doi:10.1145/3442188.3445918

[84] Andreas Hutterer and Barbara Krumay. 2024. The adoption of data spaces: Drivers toward federated data sharing. doi:10.24251/HICSS.2024.542

[85] Oana Inel, Tim Draws, and Lora Aroyo. 2023. Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11, 11 (Nov. 2023), 51–64. doi:10.1609/hcomp.v11i1.27547

[86] Umar Iqbal, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. 2023. Tracking, Profiling, and Ad Targeting in the Alexa Echo Smart Speaker Ecosystem. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. 569–583. doi:10.1145/3618257.3624803 arXiv:2204.10920 [cs].

[87] Neema Iyer. 2022. Digital Extractivism in Africa Mirrors Colonial Practices. https://hai.stanford.edu/news/neema-iyer-digital-extractivism-africa-mirrors-colonial-practices

[88] Neema Iyer, Garnett Achieng, Favour Borokini, Uri Ludger, Neema Iyer, Yahya Syabani, and Yahya Syabani. 2021. Automated Imperialism, Expansionist Dreams: Exploring Digital Extractivism in Africa. (2021). https://archive.pollicy.org/digitalextractivism/

[89] C. L. R. James. 1989. *The black Jacobins: Toussaint l'Ouverture and the San Domingo revolution* (2. ed., rev ed.). Vintage Books, a Division of Random House, Inc, New York.

[90] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. doi:10.1145/3351095.3372829

[91] Marie-Odile Junker. 2024. Data-mining and Extraction: the gold rush of AI on Indigenous Languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Sarah Moeller, Godfred Agyapong, Antti Arppe, Aditi Chaudhary, Shruti Rijhwani, Christopher Cox, Ryan Henke, Alexis Palmer, Daisy Rosenblum, and Lane Schwartz (Eds.). Association for Computational Linguistics, St. Julians, Malta, 52–57. https://aclanthology.org/2024.computel-1.8

[92] Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. arXiv:2501.18493 (Jan. 2025). doi:10.48550/arXiv.2501.18493 arXiv:2501.18493 [cs].

[93] Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 45–60.

[94] Karya. [n. d.]. Karya | We solve data needs. https://www.karya.in/

[95] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. *Forthcoming 19 Ohio St. Tech. L.J. (2023)* (2022). doi:10.2139/ssrn.4217148

[96] Hannah R. Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *Advances in Neural Information Processing Systems* 37 (Dec. 2024), 105236–105344. https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html

[97] Rob Kitchin, Juliette Davret, Carla M Kayanan, and Samuel Mutter. 2025. Assemblage theory, data systems and data ecosystems: The data assemblages of the Irish planning system. *Big Data Society* 12, 3 (2025), 20539517251352822. doi:10.1177/20539517251352822

[98] Rob Kitchin and Tracey Lauriault. 2014. Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. (2014). https://papers.ssrn.com/abstract=2474112

[99] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 100–112. doi:10.1145/3630106.3658543

[100] Naomi Klein. 2013. Naomi Klein Chats with Leanne Simpson about Idle No More. https://www.yesmagazine.org/social-justice/2013/03/06/dancing-the-world-into-being-a-conversation-with-idle-no-more-leanne-simpson

[101] Christiane Kliemann. 2020. Towards a Non-Extractive and Care-Driven Academia. https://www.developmentresearch.eu/?p=801

[102] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. arXiv:2112.01716 (Dec. 2021). doi:10.48550/arXiv.2112.01716 arXiv:2112.01716 [cs, stat].

[103] Brij Kothari. 1997. Rights to the Benefits of Research: Compensating Indigenous Peoples for their Intellectual Contribution. *Human Organization* 56, 2 (1997), 127–137. doi:10.17730/humo.56.2.j63678502x782100

[104] Sandra Kouritzin and Satoru Nakagawa. 2018. Toward a non-extractive research ethics for transcultural, translingual research: perspectives from the coloniser and the colonised. *Journal of Multilingual and Multicultural Development* 39, 8 (2018), 675–687. doi:10.1080/01434632.2018.1427755

[105] Tahu Kukutai and Donna Cormack. 2020. *"Pushing the space": Data sovereignty and self-determination in Aotearoa NZ* (1st edition ed.). Routledge, 21–35. https://www.taylorfrancis.com/reader/read-online/6abf9fc2-820b-4950-b310-eef574873fbb/chapter/pdf?context=ubx

[106] Tahu Kukutai and John Taylor. 2016. *Indigenous data sovereignty: Toward an agenda.* ANU press.

[107] Ayyüce Kızrak. 2024. What is the Data Space? Medium. https://ayyucekizrak.medium.com/what-is-the-data-space-36037d0aab2d Blog post.

[108] Veneranda Langa. 2024. Rise of Chinese surveillance tech in Africa: Development or espionage? https://www.theafricareport.com/354469/rise-of-chinese-surveillance-tech-in-africa-development-or-espionage/

[109] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. arXiv:2303.03915 (March 2023). doi:10.48550/arXiv.2303.03915 arXiv:2303.03915 [cs].

[110] Y. LeCun. [n. d.]. THE MNIST DATABASE of handwritten digits. *http://yann.lecun.com/exdb/mnist/* ([n. d.]). https://cir.nii.ac.jp/crid/1571417126193283840

[111] Peter Lee. 2024. Synthetic Data and the Future of AI. 4722162 (Feb. 2024). https://papers.ssrn.com/abstract=4722162

[112] David Leslie, Michael Katell, Mhairi Aitken, Jatinder Singh, Morgan Briggs, Rosamund Powell, Cami Rincón, Thompson Chengeta, Abeba Birhane, Antonella Perini, Smera Jayadeva, and Anjali Mazumder. 2022. Advancing Data Justice Research and Practice: An Integrated Literature Review. doi:10.5281/zenodo.6408304

[113] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. *Indigenous Protocol and Artificial Intelligence Position Paper.* Indigenous Protocol and Artificial Intelligence Working Group and

the Canadian Institute for Advanced Research, Honolulu, HI. doi:10.11573/spectrum.library.concordia.ca.00986506

[114] Jason Edward Lewis, Hēmi Whaanga, and Ceyda Yolgörmez. 2025. Abundant intelligences: placing AI within Indigenous knowledge frameworks. *AI SOCIETY* 40, 4 (April 2025), 2141–2157. doi:10.1007/s00146-024-02099-4

[115] Calvin A. Liang, Sean A. Munson, and Julie A. Kientz. 2021. Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People. *ACM Trans. Comput.-Hum. Interact.* 28, 2 (April 2021), 14:1–14:47. doi:10.1145/3443686

[116] Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative AI: open washing and the EU AI Act. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1774–1787. doi:10.1145/3630106.3659005

[117] Spencer Lilley, Gillian Oliver, Jocelyn Cranefield, and Matthew Lewellen. 2024. Māori data sovereignty: contributions to data cultures in the government sector in New Zealand. *Information, Communication Society* 27, 16 (Dec. 2024), 2801–2816. doi:10.1080/1369118X.2024.2302987

[118] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[119] LingoAI. 2025. About LingoAI | LingoAI. https://docs.lingoai.io/introduction/about-lingoai

[120] Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3933–3944. doi:10.18653/v1/2022.acl-long.272

[121] Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Manan Dey, Mohammed Hamdy, Nayan Saxena, Ahmad Mustafa Anis, Emad A. Alghamdi, Vu Minh Chien, Naana Obeng-Marnu, Da Yin, Kun Qian, Yizhi Li, Minnie Liang, An Dinh, Shrestha Mohanty, Deividas Mataciunas, Tobin South, Jianguo Zhang, Ariel N. Lee, Campbell S. Lund, Christopher Klamm, Damien Sileo, Diganta Misra, Enrico Shippole, Kevin Klyman, Lester JV Miranda, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Vipul Gupta, Vivek Sharma, Xuhui Zhou, Caiming Xiong, Luis Villa, Stella Biderman, Alex Pentland, Sara Hooker, and Jad Kabbara. 2024. Bridging the Data Provenance Gap Across Text, Speech and Video. arXiv:2412.17847 (Dec. 2024). doi:10.48550/arXiv.2412.17847 arXiv:2412.17847 [cs].

[122] LXAI. 2018. LatinX in AI (LXAI). https://www.latinxinai.org/

[123] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. 2024. "They only care to show us the wheelchair": disability representation in text-to-image AI models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3613904.3642166

[124] Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A Multidialectal Dataset of Arabic Proverbs for LLM Benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 12320–12341. doi:10.18653/v1/2025.naacl-long.613

[125] Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers. arXiv:2305.19474 (May 2023). doi:10.48550/arXiv.2305.19474 arXiv:2305.19474 [cs].

[126] Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Yuehgoh Foutse, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Samantha Dalal, et al. 2025. The World Wide recipe: A community-centred framework for fine-grained data collection and regional bias operationalisation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 246–282.

[127] Candace Maracle. 2024. How AI can help Indigenous language revitalization, and why data sovereignty is important | CBC News. *CBC News* (Aug. 2024). https://www.cbc.ca/news/indigenous/ai-indigenous-languages-1.7290740

[128] Vukosi Marivate. 2021. *Why African natural language processing now? A view from South Africa AfricaNLP*. Mapungubwe Institute for Strategic Reflection (MISTRA), 126–152. doi:10.2307/jj.12406168.11

[129] Donald Jr. Martin. 2020. Upgrading the Product Development Process to Foster Machine Learning Fairness and Ethical AI. https://www.youtube.com/watch?v=1Uyc9SPeYkA

[130] Ulises A. Mejias and Nick Couldry. 2024. *Data Grab: The New Colonialism of Big Tech and How to Fight Back*. University of Chicago Press, Chicago, IL. https://press.uchicago.edu/ucp/books/book/chicago/D/bo216184200.html

[131] Sábëlo Mhlambi and Simona Tiribelli. 2023. Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms. *Topoi* 42, 3 (2023), 867–880. doi:10.1007/s11245-022-09874-2

[132] Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–37. doi:10.1145/3555561

[133] Stefania Milan and Emiliano Treré. 2019. Big Data from the South(s): Beyond Data Universalism. *Television New Media* 20, 4 (May 2019), 319–335. doi:10.1177/1527476419837739

[134] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. doi:10.1145/3287560.3287596

[135] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy Technology* 33, 4 (Dec. 2020), 659–684. doi:10.1007/s13347-020-00405-8

[136] Cristina Jayme Montiel and Joshua Uyheng. 2022. Foundations for a decolonial big data psychology. *Journal of Social Issues* 78, 2 (2022), 278–297. doi:10.1111/josi.12439

[137] Fabio Morreale, Elham Bahmanteymouri, Brent Burmester, Andrew Chen, and Michelle Thorp. 2023. The unwitting labourer: extracting humanness in AI training. *AI SOCIETY* (May 2023). doi:10.1007/s00146-023-01692-3

[138] Luke Munn. 2024. The five tests: designing and evaluating AI according to indigenous Māori principles. *AI SOCIETY* 39, 4 (Aug. 2024), 1673–1681. doi:10.1007/s00146-023-01636-x

[139] Zachary Munn, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology* 18, 1 (Nov. 2018), 143. doi:10.1186/s12874-018-0611-x

[140] Simone Natale, Federico Biggio, Payal Arora, John Downey, Riccardo Fassone, Rafael Grohmann, Andrea Guzman, Emily Keightley, Deqiang Ji, Vincent Obia, Aleksandra Przegalinska, Usha Raman, Paola Ricaurte, and Eduardo Villanueva-Mansilla. 2025. Global AI Cultures. *Commun. ACM* 68, 9 (Aug. 2025), 37–40. doi:10.1145/3722547

[141] M. Nayebare, R. Eglash, U. Kimanuku, R. Baguma, J. Mounsey, and C. Maina. 2023. *Interim Report for Ubuntu-AI: A Bottom-up Approach to More Democratic and Equitable Training and Outcomes for Machine Learning*. San Francisco. https://generativejustice.org/uai/

[142] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2144–2160. doi:10.18653/v1/2020.findings-emnlp.195

[143] George W. Noblit and Dwight Hare. 1988. *Meta-Ethnography: Synthesizing Qualitative Studies*. https://books.google.com/books?hl=en&lr=&id=K4SuDAAAQBAJ&oi=fnd&pg=PA5&dq=noblit+and+hare+1988&ots=1g1oaCByix&sig=S30Ak-uliPXNl8RbWFBO81grNNo#v=onepage&q=noblit%20and%20hare%201988&f=false

[144] Toussaint Nothias. 2020. Access granted: Facebook's free basics in Africa. *Media, Culture Society* 42, 3 (April 2020), 329–348. doi:10.1177/0163443719890530

[145] OECD. 2024. *Explanatory memorandum on the updated OECD definition of an AI system*. Number 8 in OECD Artificial Intelligence Papers. Paris. https://doi.org/10.1787/623da898-en

[146] City of Helsinki. 2020. *AI Register*. https://ai.hel.fi/

[147] Dietmar Offenhuber. 2024. Shapes and frictions of synthetic data. *Big Data Society* 11, 2 (2024), 20539517241249390. doi:10.1177/20539517241249390

[148] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 116–126. doi:10.18653/v1/2021.mrl-1.11

[149] Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? arXiv:2311.07978 (2025). doi:10.48550/arXiv.2311.07978 arXiv:2311.07978 [cs].

[150] Chinasa T. Okolo, Kehinde Aruleba, and George Obaido. 2023. *Responsible AI in Africa—Challenges and Opportunities*. Springer International Publishing, Cham, 35–64. doi:10.1007/978-3-031-08215-3_3

[151] C Okorie and M Omino. 2024. Licensing African Datasets. https://licensingafricandatasets.com/

[152] Partnership on AI. [n. d.]. Guidance for Inclusive AI: Guidance for Inclusive AI Practicing Participatory Engagement. https://partnershiponai.org/guidance-for-inclusive-ai-new-practitioners/

[153] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Visual and Semantic Information. (2022). doi:10.48550/ARXIV.2211.15848

[154] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (March 2021), n71. doi:10.1136/bmj.n71

[155] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 881–904. doi:10.1145/3593013.3594049

[156] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 52–63. doi:10.1145/3442188.3445870

[157] Frank Pasquale and Haochen Sun. 2024. Consent and Compensation: Resolving Generative AI's Copyright Crisis. 4826695 (May 2024). doi:10.2139/ssrn.4826695

[158] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (Nov. 2021), 100336. doi:10.1016/j.patter.2021.100336

[159] Ray Pawson, Trisha Greenhalgh, Gill Harvey, and Kieran Walshe. 2005. Realist review: A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research Policy* 10, 1_suppl (July 2005), 21–34. doi:10.1258/1355819054308530

[160] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116 (2023). doi:10.48550/arXiv.2306.01116 arXiv:2306.01116 [cs].

[161] Maneesha Perera, Rajith Vidanaarachchi, Sangeetha Chandrashekeran, Melissa Kennedy, Brendan Kennedy, and Saman Halgamuge. 2025. Indigenous peoples and artificial intelligence: A systematic review and future directions. *Big Data Society* 12, 2 (2025), 20539517251349170. doi:10.1177/20539517251349170

[162] Jorge Perez. 2025. Tokenising culture: causes and consequences of cultural misalignment in large language models. https://www.adalovelaceinstitute.org/blog/cultural-misalignment-llms/

[163] Jennifer Pierre, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. 2021. Getting Ourselves Together: Data-centered participatory design research epistemic burden. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–11. doi:10.1145/3411764.3445103

[164] Claudio Pinhanez, Paulo Cavalin, Luciana Storto, Thomas Finbow, Alexander Cobbinah, Julio Nogima, Marisa Vasconcelos, Pedro Domingues, Priscila de Souza Mizukami, Nicole Grell, Majoí Gongora, and Isabel Gonçalves. 2024. Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences. arXiv:2407.12620 (2024). http://arxiv.org/abs/2407.12620 arXiv:2407.12620 [cs].

[165] Marie-Therese Png. 2023. *Paradoxes of Participation In Inclusive AI Governance: Four Key Approaches From Global South and Civil Society Discourse.* UNESCO, 269–292. https://unesdoc.unesco.org/ark:/48223/pf0000384787.locale=en

[166] QAI. 2017. Queer in AI. https://www.queerinai.com/

[167] Organizers Of QueerInAI, S Ashwin, William Agnew, Hetvi Jethwani, and Arjun Subramonian. 2021. Rebuilding Trust: Queer in AI Approach to Artificial Intelligence Risk Management. *Queer in AI Workshop at NeurIPS 2021* (2021). queerinai.org/risk-management

[168] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1882–1895. doi:10.1145/3593013.3594134

[169] Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessico Ojo, Atnafu Lambebo Tonja, Maushami Chetty, Wilhelmina NdapewaOnyothi Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2025. The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages. arXiv:2502.15916 (2025). doi:10.48550/arXiv.2502.15916 arXiv:2502.15916 [cs].

[170] Inioluwa Deborah Raji and Joy Buolamwini. 2023. Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Commun. ACM* 66, 1 (Jan. 2023), 101–108. doi:10.1145/3571151

[171] Qazi Mamunur Rashid, Erin van Liemt, Tiffany Shih, Amber Ebinama, Karla Barrios Ramos, Madhurima Maji, Aishwarya Verma, Charu Kalia, Jamila Smith-Loud, Joyce Nakatumba-Nabende, Rehema Baguma, Andrew Katumba, Chodrine Mutebi, Jagen Marvin, Eric Peter Wairagala, Mugizi Bruce, Peter Oketta, Lawrence Nderu, Obichi Obiajunwa, Abigail Oppong, Michael Zimba, and Data Authors. 2025. Amplify Initiative: Building A Localized Data Platform for Globalized AI. arXiv:2504.14105 (April 2025). doi:10.48550/arXiv.2504.14105 arXiv:2504.14105 [cs].

[172] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. [n. d.]. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *2020* ([n. d.]).

[173] Mohammad Rashidujjaman Rifat, Abdullah Hasan Safir, Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohammad Ruhul Amin, and Syed Ishtiaque Ahmed. 2024. Data, Annotation, and Meaning-Making: The Politics of Categorization in Annotating a Dataset of Faith-based Communal Violence. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2148–2156. doi:10.1145/3630106.3659030

[174] Katja Rogers, Teresa Hirzle, Sukran Karaosmanoglu, Paula Toledo Palomino, Ekaterina Durmanova, Seiji Isotani, and Lennart E. Nacke. 2024. An Umbrella Review of Reporting Quality in CHI Systematic Reviews: Guiding Questions and Best Practices for HCI. *ACM Trans. Comput.-Hum. Interact.* 31, 5 (Nov. 2024), 57:1–57:55. doi:10.1145/3685266

[175] Caroline Running Wolf and Noelani Arista. 2020. *Indigenous Protocols in Action.* Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research, Honolulu, HI, 93–101. doi:10.11573/spectrum.library.concordia.ca.00986506

[176] Amanda Sabin. 2024. Prioritizing Indigenous Participation and Compensation in Research. *Journal of Critical Global Issues* 1, 1 (Feb. 2024). doi:10.62895/2997-0083.1004

[177] Jathan Sadowski. 2019. When data is capital: Datafication, accumulation, and extraction. *Big Data Society* (Jan. 2019). doi:10.1177/2053951718820549

[178] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. arXiv:2101.09995 (Jan. 2021). http://arxiv.org/abs/2101.09995 arXiv:2101.09995 [cs].

[179] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. doi:10.1145/3411764.3445518

[180] Bonaventure Saturday and Bonnita Nyamwire. 2023. *Towards Effective Data Governance in Africa (Progress, Initiatives and Challenges).* https://pollicy.org/resource/towards-effective-data-governance-in-africa-progress-initiatives-and-challenges/

[181] Morgan Klaus Scheuerman. 2024. In the walled garden: Challenges and opportunities for research on the practices of the AI tech industry. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 456–466.

[182] Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. 2023. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–33. doi:10.1145/3579488

[183] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. (2022). doi:10.48550/ARXIV.2210.08402

[184] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. (2021). doi:10.48550/ARXIV.2111.02114

[185] Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 724–731. doi:10.18653/v1/2022.acl-short.82

[186] Eve Kosofsky Sedgwick. 1990. *Epistemology of the Closet.* University of California Press. Google-Books-ID: u5jgaOhhmpgC.

[187] Ali Akbar Septiandri, Marios Constantinides, and Daniele Quercia. 2024. WEIRD ICWSM: How Western, Educated, Industrialized, Rich, and Democratic is Social Computing Research? arXiv:2406.02090 (2024). doi:10.48550/arXiv.2406.02090 arXiv:2406.02090 [cs].

[188] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. 160–171. doi:10.1145/3593013.3593985 arXiv:2305.06415 [cs].

[189] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. arXiv:2210.05791 (2023). http://arxiv.org/abs/2210.05791 arXiv:2210.05791 [cs].

[190] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 642–659. doi:10.1145/3630106.3658931

[191] Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619* (2024).

[192] Linda Tuhiwai Smith. 2021. *Decolonizing Methodologies: Research and Indigenous Peoples* (third edition ed.). Bloomsbury Publishing. Google-Books-ID: EwA1EAAAQBAJ.

[193] Aivin V Solatorio, Gabriel Stefanini Vicente, Holly Krambeck, and Olivier Dupriez. 2024. Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers. *arXiv preprint arXiv:2410.10665* (2024).

[194] Gayatri Chakravorty Spivak. 1994. *Can the Subaltern Speak?* Routledge, London, 66–111.

[195] Spyros Spyrou. 2024. From extractivist practices and the child-as-data to an ethics of reciprocity and mutuality in empirical childhood research. *Childhood* 31, 1 (Feb. 2024), 3–12. doi:10.1177/09075682231220158

[196] Yolande Strengers, Jathan Sadowski, Zhuying Li, Anna Shimshak, and Florian "Floyd" Mueller. 2021. What Can HCI Learn from Sexual Consent?: A Feminist Process of Embodied Consent for Interactions with Emerging Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3411764.3445107

[197] Harini Suresh and John V. Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9. doi:10.1145/3465416.3483305 arXiv:1901.10002 [cs, stat].

[198] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 667–678. doi:10.1145/3531146.3533132

[199] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1609–1621.

[200] Alex S. Taylor. 2011. Out there. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 685–694. doi:10.1145/1978942.1979042

[201] Jordan Taylor, Wesley Hanwen Deng, Kenneth Holstein, Sarah Fox, and Haiyi Zhu. 2024. Carefully Unmaking the "Marginalized User:" A Diffractive Analysis of a Gay Online Community. *ACM Transactions on Computer-Human Interaction* (2024), 3673229. doi:10.1145/3673229

[202] Linnet Taylor. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data Society* 4, 2 (Dec. 2017), 205395171773633. doi:10.1177/2053951717736335

[203] Te Hiku Media. [n. d.]. *Kaitiakitanga License.* https://github.com/TeHikuMedia/Kaitiakitanga-License GitHub repository.

[204] Jaime A. Teixeira da Silva. 2022. Handling Ethics Dumping and Neo-Colonial Research: From the Laboratory to the Academic Literature. *Journal of Bioethical Inquiry* 19, 3 (2022), 433–443. doi:10.1007/s11673-022-10191-x

[205] Tesfa-Alem Tekle. 2020. Refugees in Ethiopia's camps raise privacy and exclusion concerns over UNHCR's new digital registration. https://globalvoices.org/2020/03/19/refugees-in-ethiopias-camps-raise-privacy-and-exclusion-concerns-over-unhcrs-new-digital-registration/

[206] Jim Thatcher, David O'Sullivan, and Dillon Mahmoudi. 2016. Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space* 34, 6 (Dec. 2016), 990–1006. doi:10.1177/0263775816633195

[207] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-first Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3441852.3471225

[208] Scott Timcke. 2024. AI and the digital scramble for Africa. https://roape.net/2024/07/11/ai-and-the-digital-scramble-for-africa/

[209] Christina Timko, Malte Niederstadt, Naman Goel, and Boi Faltings. 2023. Incentive Mechanism Design for Responsible Data Governance: A Large-scale Field Experiment. *J. Data and Information Quality* 15, 2 (2023), 16:1–16:18. doi:10.1145/3592617

[210] Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2024. InkubaLM: A small language model for low-resource African languages. arXiv:2408.17024 (2024). doi:10.48550/arXiv.2408.17024 arXiv:2408.17024 [cs].

[211] Paola Tubaro, Antonio A. Casilli, Maxime Cornet, Clément Le Ludec, and Juana Torres Cierpe. 2025. Where does AI come from? A global case study across Europe, Africa, and Latin America. (Feb. 2025). doi:10.1080/13563467.2025.2462137 arXiv:2502.04860 [cs].

[212] Mozilla Common Voice. [n. d.]. Mozilla Common Voice. https://commonvoice.mozilla.org/

[213] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 214–229. doi:10.1145/3531146.3533088

[214] Cedric Deslandes Whitney and Justin Norman. 2024. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1733–1744. doi:10.1145/3630106.3659002

[215] David Gray Widder, Meredith Whittaker, and Sarah Myers West. 2024. Why 'open' AI systems are actually closed, and why this matters. *Nature* 635, 8040 (Nov. 2024), 827–833. doi:10.1038/s41586-024-08141-1

[216] Tanja Wiehn. 2024. Synthetic Data: From Data Scarcity to Data Pollution. *Surveillance Society* 22, 4 (Dec. 2024). doi:10.24908/ss.v22i4.18327

[217] Julia Carrie Wong. 2019. The Cambridge Analytica scandal changed the world – but it didn't change Facebook. *The Guardian* (March 2019). https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook

[218] Yung-Hsuan Wu. 2024. Capturing the unobservable in AI development: proposal to account for AI developer practices with ethnographic audit trails (EATs). *AI and Ethics* (2024). doi:10.1007/s43681-024-00535-1

[219] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. 2018. A Nutritional Label for Rankings. In *Proceedings of the 2018 International Conference on Management of Data*. 1773–1776. doi:10.1145/3183713.3193568 arXiv:1804.07890 [cs].

[220] Edibe Betul Yucer. 2025. AI is finally trying to speak African languages. Will this end a historic neglect? *TRT Global* (Aug. 2025). https://trt.global/afrika-english/article/359e1362af39

[221] Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 1529–1541. doi:10.18653/v1/2022.acl-long.108

[222] Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding Dialect Disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3701–3720. doi:10.18653/v1/2022.acl-long.258

A Corpus Dataset

Upon acceptance, we will release the full coded corpus as a versioned repository on GitHub. The repository will include:

- **Corpus data**: bibliographic metadata, coding assignments (orientation, pipeline stage, historical era, geographic focus, triangle coverage), and structured summaries.
- **Versioning**: DOI-tagged releases via Zenodo to support citation and reproducibility.
- **Community feedback**: issues and pull requests for corrections, additions, and reclassifications, reviewed by maintainers.
- **Multilingual expansion**: framework accommodates sources in additional languages as contributors extend coverage.
- **Transparency**: coding definitions (C) will function as a living codebook, with archived rationales for inclusion and coding decisions.

We plan the repository to function as a living corpus, open to contestation and expansion. Our aim is a transparent, multilingual dataset that supports sustained inquiry into AI data production and its impacts on Indigenous, underrepresented, and underserved communities worldwide. This open approach addresses the documented need for reproducible, community-accessible resources in HCI and critical computing research, in addition to other fields.

B Search Strategy and Corpus Construction

B.1 Detailed Inclusion and Exclusion Criteria

Table 5 outlines the criteria guiding corpus construction. Primary and expanded criteria required sources to address AI contexts, data production practices, and/or community impacts. Specific inclusion standards emphasized substantive engagement, analytical connectivity, and relevance to power dynamics, while exclusions removed works lacking AI context, community focus, or adequate treatment of data production.

B.2 ACM Digital Library Search Strategy

Database searches were conducted iteratively between August 2024 and January 2025, complementing network referrals and citation snowballing. The final structured ACM Digital Library search was executed on January 31, 2025, using the advanced search interface with abstract and full-text indexing via personal subscription.

Table 6 reports the four primary ACM query sets and their outcomes.

Aggregate results from the January 2025 ACM searches are summarized in Table 8. Across 1,914 hits, 1,201 items were screened, yielding 153 that met criteria and 48 unique sources after full-text review and duplicate removal.

Table 5. Detailed inclusion and exclusion criteria for corpus construction.

Category	Criteria
Primary criteria (sources addressing all three areas)	– AI contexts – Data production practices (sourcing, processing, governance, infrastructure) – Community impacts (Indigenous, underrepresented, underserved communities worldwide)
Expanded criteria (sources addressing at least two areas)	– Extractive data production practices in AI contexts – Less-extractive practices involving Indigenous, underrepresented, underserved communities – Impacts of AI systems on these communities worldwide – Foundational critical theory and historical works on power in knowledge production
Specific inclusion standards	– Substantive engagement: analysis beyond cursory mention – Analytical connectivity: insights linking AI, data practices, and community outcomes – Relevance to power dynamics: addressing equity, extraction, or agency
Exclusions	– Community-Based Research (CBR) studies without AI context – AI ethics discussions without data practice or community focus – Case studies with minimal data production analysis – Non-English materials (pragmatic limitation acknowledged)

Table 6. ACM Digital Library search queries and results.

Query	Exact search string	Results screened	Potentially relevant
Q1	(("data collection" OR "data production" OR "data curation" OR "dataset development") AND ("artificial intelligence" OR "machine learning" OR "AI") AND ("marginalized" OR "underrepresented" OR "underserved" OR "community" OR "indigenous"))	51 abstracts + 300 full-texts	45 (7 abstracts, 38 full-texts)
Q2	(("extractive" OR "exploitative" OR "data colonialism") AND ("data practices" OR "dataset construction") AND ("communities" OR "workers" OR "labor"))	3 abstracts + 182 full-texts	13 (1 abstract, 12 full-texts)
Q3	(("crowdsourcing" OR "platform labor") AND ("bias" OR "fairness" OR "ethics") AND ("marginalized" OR "vulnerable populations" OR "community harm"))	491 full-texts (200 screened)	32
Q4	(("participatory design" OR "community-led" OR "co-design") AND ("ai development" OR "dataset creation") AND ("sovereignty" OR "community engagement" OR "ethical data"))	122 full-texts	12

B.3 Additional Database Searches

Similar comprehensive search strategies were applied to IEEE Xplore, ScienceDirect, Taylor & Francis Online, Wiley Online Library, Google Scholar, and Springer Link, following the same phased approach for queries.

Table 7. Targeted ACM venue-specific searches.

Venue	Exact search string	Venue filter	Results summary
CHI Conference Proceedings	(("data collection" OR "data production" OR "data curation" OR "dataset development") AND ("artificial intelligence" OR "machine learning" OR "AI") AND ("marginalized" OR "underrepresented" OR "underserved" OR "community" OR "indigenous"))	CHI Conference on Human Factors in Computing Systems (all years)	296 hits; screened: first 200; potentially relevant: 15
FAccT Proceedings	(("data collection" OR "data production" OR "data curation" OR "dataset development") AND ("artificial intelligence" OR "machine learning" OR "AI") AND ("marginalized" OR "underrepresented" OR "underserved" OR "community" OR "indigenous"))	ACM Conference on Fairness, Accountability, and Transparency	143 hits; screened: all; potentially relevant: 37

Table 8. ACM search results summary.

Category	Count
Total primary searches (query sets)	6
Total venue-specific searches	2
Total hits across all searches	1,914
Total items screened (varied by search size)	1,201
Items meeting inclusion criteria after screening	153
Items retained after full-text review	89
Final unique sources for corpus (after duplicate removal)	48

B.4 Network-Based Source Discovery

Table 9 summarizes how community organizations appear in the corpus, distinguishing between direct data production outputs, project citations, and broader ecosystem references.

B.4.1 Citation Snowballing Protocol.

- Seed papers: 20 foundational works from initial database searches
- Forward citations: Google Scholar alerts through February 2025
- Backward citations: Systematic reference list review

B.5 Quality Assessment and Coding Procedures

White literature sources (258 sources, 74% of corpus) underwent standard peer review processes. Grey literature (92 sources, 26% of corpus) consisted of organizational reports, policy documents, and community outputs from established organizations and individuals rather than informal publications.

Table 9. Community organizations in corpus sourcing.

Category	Organizations
Analyzed (general)	Masakhane; Black in AI; Queer in AI; Indigenous Protocol & AI Collective / Abundant Intelligences; Ghana NLP; Sunbird AI; Lelapa AI; LingoAI; Karya; Vaani; LatinX in AI
Analyzed with direct data production outputs	Masakhane; Ghana NLP; Sunbird AI; Lelapa AI; LingoAI; Karya; Vaani
Cited projects (community-led but represented via publications, not organizational entries)	Queer in AI; Indigenous Protocol & AI Collective / Abundant Intelligences
Ecosystem collectives (referenced in discussion, not coded)	Black in AI; LatinX in AI

Inter-coder reliability was established through collaborative consensus rather than statistical measures. Each co-author independently identified foundational sources, which were pooled to form the initial corpus. The first author then assumed primary responsibility with team oversight through two additional consensus rounds (August–September 2024). In each round, the first author presented approximately 100 coded sources for collective team review. Disagreements on orientation, triangle coverage, and pattern assignment were resolved through discussion until consensus was reached. This process established shared frameworks before the first author completed the full 350-source corpus.

B.6 Search Result Summary

Table 10 summarizes how the 350 corpus sources were identified across different discovery pathways.

Table 10. Discovery method distribution (N=350 sources).

Method	Sources	Percent
Database searches	175	50%
Existing networks/organizations	72	21%
Citation snowballing	51	15%
Iterative keyword search	31	9%
Hand-searching journals	21	6%
Total	**350**	**100%**

C Coding Schema and Corpus Composition

C.1 Overview

3.3 introduces the coding schema applied to all sources in the corpus. Section C.2 below provides full definitions for each column, followed by examples of triangle coverage coding (Table 12) and a corpus composition summary (C.4).

C.2 Coding Definitions

Column	Principles
Identifier	In-line APA citation (author surname and year) used as a unique ID for tracking within the corpus.
APA Citation	Full APA reference for the source.
Title	Title of the publication or output.
Orientation	One of three values: • Extractive: undermines consent, compensation, or benefit. • Principles less-extractive: normative frameworks promoting stewardship, sovereignty, accountability. • Practices less-extractive: operationalized, community-led, participatory, or sovereignty-based initiatives.
Domain Categories	Thematic focus of the source (controlled list, multiple possible): • Community impacts and relations • Critical theory • Data labor • Data practices • Ethics frameworks
Pipeline Stage	Specific process within a pipeline stage (controlled list): • Problem Understanding and Formulation • ML System Design and Development • Deployment and Impact • Cross-pipeline

Column	Principles
Pipeline Sub-stage	Specific process within a pipeline stage (controlled list): • Institutional Prioritization and Funding • Product Conception and Design • Data Selection, Collection and Annotation • Model Architecture Selection and Design • Model Training and Evaluation • Product Testing • Product Launch • Cross-pipeline
Historical Era	Era of data production practice (controlled list): • Era 1: Curated datasets (pre-2009); no sources in corpus • Era 2: Crowdsourced benchmarks (2009–2017) • Era 3: Web-scraped/foundation models (2017–present) • Multi-era: Spans multiple eras or provides historical analysis
Primary Pattern(s) / Pathway(s)	The specific extractive mechanism or less-extractive alternative identified in findings (see Tables 1–2 in main text). Between one and three tags were assigned per source in order of relevance. For sources that provide conceptual, historical, or framing contributions without mapping directly onto one of the synthesized patterns or pathways, we assigned Other/NA (conceptual framing).
Triangle Coverage	Vertices of analytic triangle substantively engaged: AI contexts (A), data production practices (D), community impacts (C). Codes reflect combinations (AD, AC, DC, ADC) or single focus (C only). Engagement must exceed passing mention. See B.4.
How Source Was Found	White literature (journal papers, conference proceedings, books) or Grey literature (reports, policy documents, theses, community outputs, blogs).
Keywords	3–5 terms for coding/search, ordered Geography → Data/technical → Community/impact.

Column	Principles
Geographic Region of Focus	Region or community under study (controlled list): Africa, APAC, EU/UK, LatAm, MENA, North America, Oceania, Multiple regions, Not regionally specific (globally framed advocacy, transnational collectives, or technical works not tied to one region).
Author Affiliations	High-level institutional grouping of authors. If multiple affiliations, code majority grouping here; record full details in Authorship & Positionality Context. Controlled list: Academic; Government; Industry; NGO/Non-profit; Mixed; Journalist/Other/Not sure
Geographic Area of Author(s) Institution	Full institution name and country of the lead author(s) Region of lead author's institution (controlled list, same regions as above).
Authorship and Positionality Context	Complete authorship profile, including all institutions, geographic distribution, equal contribution notes, and any relevant statements on positionality or disciplinary traditions.
Summary	≤ 120-word synopsis. Structure: Topic → Method → Findings → Link to AI data production + community impacts.

C.3 Triangle Coverage

Triangle coverage indicates which vertices of the analytic triangle are substantively engaged: AI systems/contexts (A), data production practices (D), and community impacts (C). Codes reflect combinations (AD, AC, DC, ADC) or single focus (C). Engagement must exceed passing mention.

C.4 Corpus Composition Summary

Table 13 summarizes the composition of our 350-source corpus across orientation, source type, geographic focus, author affiliation, pipeline stage, and historical era.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 12. Triangle coverage coding definitions showing how corpus sources engage AI systems (A), data production practices (D), and community impacts (C)

Coverage	Description	Example Identifiers
ADC	Engages AI systems AND data sourcing, processing, governance, infrastructure, AND community impacts together	Duncan et al. [60], Rifat et al. [173], Running Wolf and Arista [175]
DC	Links data sourcing, processing, governance, infrastructure to community outcomes, but does not extensively analyze AI systems	Eglash et al. [63], Kukutai and Cormack [105], Lilley et al. [117]
AC	Connects AI system behavior or deployment to community impacts, but does not extensively engage data sourcing, processing, governance, infrastructure	Asiedu et al. [18], Birhane [28], Shelby et al. [189]
AD	Examines AI systems in relation to data sourcing, processing, governance, infrastructure, but does not extensively analyze community impacts	García et al. [70], Longpre et al. [121], Suresh and Guttag [197]
C	Focuses on community experiences or governance alone, but does not extensively analyze AI systems or data sourcing, processing, infrastructure	Durante et al. [61], Fricker [68], James [89]

Table 13. Corpus composition summary (N=350 sources)

Category	Sub-category	Count (%)
Orientation	Extractive practices	142 (41%)
	Principles less-extractive	115 (33%)
	Practices less-extractive	93 (27%)
Source Type	White literature	258 (74%)
	Grey literature	92 (26%)
Geographic Focus	Not regionally specific	151 (43%)
	Multiple areas	61 (17%)
	North America	40 (11%)
	Africa	38 (11%)
	Oceania	19 (5%)
	APAC	15 (4%)
	EU/UK	11 (3%)
	LatAm	12 (3%)
	MENA	3 (1%)
Author Affiliation (lead only)	Academic	182 (52%)
	Mixed	96 (27%)
	Industry	37 (11%)
	NGO/Non-profit	20 (6%)
	Journalist/Other	11 (3%)
	Government	4 (1%)
Pipeline Stage	ML System Design & Development	183 (52%)
	Problem Understanding & Formulation	90 (25%)
	Cross-pipeline	55 (16%)
	Deployment & Impact	22 (6%)
Historical Era	Era 3 (2017–present)	241 (69%)
	Multi-era	95 (27%)
	Era 2 (2009–2017)	14 (4%)
	Era 1 (pre-2009)	0 (0%)