Understanding AI Data Production & Community Impacts Worldwide: A Multivocal Literature Review

ANONYMOUS AUTHOR(S)

Artificial intelligence (AI) depends on data production: the sociotechnical process that transforms human knowledge into computational resources. The consequences of these processes fall disproportionately on Indigenous, underrepresented, and underserved communities, yet the connections among AI systems, data practices, and community impacts have not been systematically examined. We conduct a Multivocal Literature Review (MLR) integrating 350 academic and grey-literature sources to analyze how AI systems, data practices, and community impacts intersect. Across five analytic domains—Data Relations, Data Labor, Data Representation, Data Infrastructure, and Data Governance—we distinguish extractive data production mechanisms that prioritize scale, opacity, and labor externalization from high-agency pathways in which communities exercise authority. We contribute (1) a multivocal review that positions data production as a site of sociotechnical power rather than a technical prerequisite; (2) implications for responsible computing research including upstream infrastructure as a design site, provenance-first architectures, and federated data governance supporting community sovereignty; (3) methodological illustration of multivocal synthesis for bridging academic research with practitioner knowledge; and (4) an open corpus mapping sources across pipeline stages, historical eras, and geographic contexts.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Social and professional topics →** *Computing / technology policy*; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: Human-centred computing, AI, ML pipeline, data production, extractive practices, underserved communities, Indigenous data sovereignty, data collection

ACM Reference Format:
Anonymous Author(s). 2018. Understanding AI Data Production & Community Impacts Worldwide: A Multivocal Literature Review. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 26 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Contemporary artificial intelligence (AI) systems depend on data. As approaches have advanced over the past three decades, the scale and composition of data needs have transformed: from small expert-curated datasets like MNIST [81], to massive crowdsourced benchmarks such as ImageNet [37], and now to foundation models trained on billions of scraped web documents, images, and interaction traces [124, 136]. Although scale reduces some sampling limitations, web-scraped corpora inherit the biases of who publishes online, what content platforms permit, and which languages dominate digital spaces. The opacity and complexity of the machine learning (ML) pipeline [21], as well as the diversity and amount of human knowledge and labor needed [122, 158], have expanded dramatically. These developments span both discriminative systems (designed for prediction and classification) and generative systems (designed to produce text, images, or other content). Public attention has shifted toward generative AI since 2022, but the data production practices underlying both approaches share the foundational concerns examined here.

Manuscript submitted to ACM

The historical trajectory matters because the choices made in gathering and curating data directly shape which communities benefit from AI systems and which communities bear their costs. Data is made, not found. It is produced through a series of choices about what to gather, how to curate it, and under what terms. Decisions made upstream and throughout the pipeline can either create or mitigate harms, which fall disproportionately on underserved, underrepresented, and Indigenous communities [5, 12, 75, 85, 98, 118]. Facial recognition systems trained on demographically skewed datasets misidentify dark-skinned faces at higher rates, contributing to wrongful arrests [20, 119]; language models trained on web text reproduce stereotypes [46] and fail to serve speakers of low-resource languages [116]; and biometric data collected from refugees without meaningful consent enables surveillance infrastructures that follow displaced populations across borders [63, 161]. Despite growing critical attention to algorithmic harms and dataset bias, data production itself—a sociotechnical process through which human knowledge becomes computational input—has rarely been treated as a central object of inquiry.

Viewed through the lens of responsible computing, upstream data practices are sites where accountability, consent, and community benefit are negotiated or circumvented. Our approach leverages Critical Computing as a diagnostic lens and Social Justice as a normative orientation. Critical Computing offers diagnostic tools for analyzing how data practices reflect institutional priorities and labor arrangements rather than objective "ground truth" modeled by engineers. Social Justice complements the diagnosis by asking how data work might redistribute agency, benefit, and governance toward the communities whose knowledge and labor support AI systems. Together, the two perspectives clarify why upstream data production warrants sustained attention from researchers concerned with the societal impacts of computing.

Relevant literature spans disciplines and publication ecosystems. Human–computer interaction (HCI) contributes a long-standing body of research on how sociotechnical systems enact power, from canonical postcolonial critiques [62, 149] to recent work analyzing "extractive" dynamics in ICT4D research [44], and epistemic injustice (i.e., the systematic devaluation of certain groups' knowledge, testimony, and interpretive frameworks [4, 164]. Review literature in HCI and adjacent fields offers insight into how AI systems affect communities: Shelby et al. [141] map harms experienced by underserved groups; Wang et al. [160] synthesize findings across disability contexts; [130] examine context mismatches in AI deployment. Complementary work traces how collection and annotation practices embed exclusions in datasets [38], identifies structural gaps in AI ethics scholarship [14, 16], analyzes how misabstraction cascades through sociotechnical systems [34]; and takes stock of social-justice commitments within HCI [25].

In aggregate, the literature illuminates important dimensions of a three-part relationship between AI systems (the models and pipelines that process data), data production practices (the sourcing, annotation, and curation decisions that shape training corpora), and community impacts (the consequences—beneficial or harmful—for the populations whose knowledge, labor, or lives are represented in or affected by these systems). Yet, most scholarship examines one or two dimensions rather than synthesizing across all three. Most reviews also draw primarily on academic publications, capturing the scientific state-of-the art but leaving less visible documentation of the state-of-practice. A synthesis that spans the full relationship and draws on evidence from both academic and practitioner sources is needed, as is a method suited to fragmented evidence landscapes. Multivocal literature review (MLR) offers such an approach, synthesizing knowledge that circulates across publication ecosystems by integrating white literature (peer-reviewed academic publications) with grey literature (policy reports, organizational materials, community outputs) [49]. We adopt MLR to assemble a corpus of 350 sources examining AI data production and its impacts on underserved, underrepresented, and Indigenous communities.

We treat *data production* as the complex sociotechnical process through which data is defined, gathered, curated, and controlled across model pipelines.[1] The production framing supports a move beyond *data collection* as a routine methodological disclosure or neutral technical artifact, instead centering the institutional choices, power relations, and consequences that underpin AI development and determine who benefits from or bears its costs.

Building on STS and HCI scholarship that has long recognized data production as a sociotechnical site where power is negotiated, our synthesis documents and catalogs this dispersed literature and identifies specific mechanisms across the ML pipeline. We discuss both extractive practices that centralize control and high-agency alternatives through which communities exercise authority. In summary, we contribute:

- A multivocal literature review synthesizing 350 sources across academic and grey literature, organized through five analytic domains—Data Relations, Data Labor, Data Representation, Data Infrastructure, and Data Governance—that highlight where AI systems, data production practices, and community impacts intersect;
- Opportunities for research and practice in responsible computing, including upstream data infrastructure as a design site, provenance-first architectures, federated learning for community sovereignty, and ethics review paradigms that scrutinize data production;
- Methodological illustration of multivocal synthesis for sociotechnical inquiry, documenting how grey literature surfaces practitioner knowledge and regionally-grounded perspectives underrepresented in academic venues;
- An open corpus of 350 sampled sources mapped across pipeline stages, historical eras, and geographic contexts, with structured summaries and documented rationales, publicly available at [URL removed for review]

1.1 Key Terms and Definitions

1.1.1 Artificial Intelligence. For clarity, we use "AI" in this paper primarily in its modern ML sense, as a system that learns patterns and rules from training data to create a predictive model [21]. The resulting model must then be evaluated for reliability and generalization using a separate, independent dataset for testing [90]. As Paullada et al. [115] document, such systems depend on datasets whose construction involves consequential choices about sourcing, annotation, and evaluation—choices that have received insufficient critical scrutiny.

We contextualize our discussion of AI within its broader historical trajectory of research and development but focus on the current statistical and data-driven era of AI that facilitates many contemporary extractive regimes [52]. We intend our discussion to be situated not merely as a critique of modern ML but as a reflection on a continuous thread within technological and social history.

1.1.2 Extractive. We use "extractive" in this paper to denote high-asymmetry or dispossessive practices, building upon its conventional association with Indigenous marginalization and digital forms of resource appropriation. This definition is designed to encompass diverse historical and contemporary manifestations of power imbalances that result in one party's advantage at the expense of another's autonomy or resources [17, 110]. Illustrative examples of such high-asymmetry practices are evident in historical contexts, such as the exploitative labor practices of UK coal mining [32]; the profoundly unethical nature of the Tuskegee syphilis experiment in the United States and the untreated carcinoma study in New Zealand [67, 114]; and contemporary issues like large-scale industrial mining [104] and pervasive AI surveillance [112]. By broadening this definition, our objective is to more accurately encapsulate the systemic character of extraction across various domains. In contrast, we describe as high-agency examples of

[1] We share with Miceli & Posada [94] an emphasis on "production" to foreground relations of power and knowledge in data and labor, which echoes the "assemblage" approach of Kitchin et al. [73], also rooted in Foucauldian critique.

principles and practices in the literature which prioritize active participation, equitable distribution of power, and community-defined obligations [120, 167]. These examples appear in contexts where communities, practitioners, or institutions negotiate shared authority, shape the terms of data contribution, or establish governance arrangements that align data use with locally grounded priorities.

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 [86, 141]. 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" [127]. 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 [150].

2 Background

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 [45, 145] and "situated knowledges" [58] interrogate how knowledge systems encode relations of domination. Historical analyses of colonial resistance show alternative epistemologies and organizing strategies [65].

Black feminist theory articulates intersectional approaches to structural power, from early collective statements [28] to analyses of interlocking systems of oppression [29]. Gender and queer theory establish frameworks for analyzing the production of normativity [22], binary logics [139], and classificatory power [27]. Indigenous studies center community sovereignty and relational ethics [8, 36, 84] and provide frameworks for decolonizing knowledge production in research [144] and AI data practices [18]. Critical data studies crystallize a complementary set of concerns for the digital context, with a focus on datafication, surveillance, and governance [74].

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 [38]. As demands for larger models have intensified, practices have shifted from small, carefully curated corpora, to large datasets assembled through web-scraping and crowdsourced annotations, 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 [80]. Others, such as the UCI Adult dataset [41] derived from census records for income prediction, embedded gendered and racial assumptions in their feature design. Even at this small scale, choices about inclusion and categorization reflected institutional priorities with uneven consequences for the populations described.

Era 2. Large curated datasets expanded scale through crowdsourced annotation of web-scraped content, exemplified by ImageNet [37] and MS COCO [88]. This era accelerated deep learning [35, 53] but shifted labor from domain experts to distributed workers, often in Global Majority regions, as well as automated web-scraping efforts, ultimately emphasizing 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 [7, 30] and C4 [42, 124], LAION [136, 137], Refined Web [117], and ClueWeb22 [109] are assembled through automated scraping at an unprecedented scale. Such production efforts shape and are shaped by competitive foundation model development [13, 148], spanning both discriminative systems and the generative models that have drawn heightened public attention since 2022. Alongside and often in response, smaller, highly curated datasets emerged, produced through participatory methods and community partnerships. Examples include ROOTS [79], Masakhane's African language collections [106], and Cohere's multilingual Aya Dataset [143].

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 [60]. 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 [87, 165].

The three eras feature distinct technical capabilities, institutional arrangements, and labor configurations. Each era introduced new mechanisms through which extractive patterns could scale, while also producing the conditions under which communities organized responses. While both Era 2 and Era 3 are founded in web-scraped datasets, the scale at which Era 3 extracts data is unprecedented. As such, 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.

3 Methodology

We conducted an MLR to assemble a structured body of evidence on AI data production and its impacts on underserved, underrepresented, and Indigenous communities. We treated academic and grey sources as complementary evidence streams. The research-design diagram and accompanying materials are available on the companion site: [URL removed for review]. Supplementary materials—including complete search queries, screening logs, datasheet field definitions, and coding examples—are available at the same location.

Because our inquiry spans three interrelated dimensions—AI systems, data production practices, and community impacts—we employed a tripartite scoping approach (A/D/C) to ensure comprehensive coverage. This scoping approach defined the boundaries of our review: all sources engage substantively with community impacts (C), while varying in how directly they address AI systems (A) and data production practices (D). All sources engage substantively with community experiences, power dynamics, or consequences in contexts relevant to AI data production. Sources varied in whether they directly addressed AI systems (A) and data production practices (D), or provided foundational understanding that informs interpretation of these dimensions. Section 3.2 provides more detail on how we created the corpus based on this scoping approach.

3.1 Search Strategy

We developed the search strategy by deriving keywords from the A/D/C scoping approach. We searched seven academic databases between August 2024 and January 2025: ACM Digital Library, IEEE Xplore, ScienceDirect, Taylor & Francis

Online, Wiley Online Library, Springer Link, and Google Scholar. We iteratively developed boolean search strings with AND/OR terms across variants of A/D/C terms using Boolean operators. Titles and abstracts were screened first, followed by full-text assessment for items meeting initial criteria. Table 1 shows the key and supplementary search terms of our inquiry.

Table 1. Key and Supplementary Search Terms

Dimension	Key Terms	Supplementary Terms
AI Systems (A)	Artificial Intelligence, Machine Learning, AI, ML	Large Language Models, LLMs, Computer Vision, Foundation Models, Neural Networks, Automated Systems, Algorithmic Systems, Deep Learning
Data Production (D)	Data Collection, Data Production, Dataset Creation, Data Curation, Data Practices	Annotation, Labeling, Data Labor, Crowdsourcing, Web Scraping, Data Extraction, Dataset Development, Data Work, Data Gathering, Responsible AI
Community Impacts (C)	Indigenous, Marginalized, Underrepresented, Underserved, Community	Global Majority, Global South, Data Sovereignty, Linguistic Diversity, Cultural Context, Extraction, Appropriation, Bias, Fairness, Harm, Safety

We developed four primary query sets: foundational (targeting core data production practices in AI contexts affecting communities), extraction frame (targeting exploitative practices), data labor (focusing on crowdsourcing and platform labor), and alternatives (seeking participatory and community-led approaches). The ACM Digital Library search illustrates our results. Across the four query sets, 1,914 hits yielded 1,201 items screened, 153 meeting initial criteria, and 48 unique sources after full-text review and duplicate removal. Similar strategies applied to the remaining six databases. Database searches contributed 174 sources, representing 50% of the final corpus. See supplementary materials for more search details.

For grey literature we used different methods. Following Garousi et al. [49], we used general Google Search and systematically examined organizational ecosystems engaged in AI data work, prioritizing organizational reports, policy documents, and community outputs from established entities. Three complementary methods supplemented database and grey literature searches. Citation snowballing [166] from 20 seed papers tracked forward and backward citations iteratively, contributing 51 sources (15%). Hand-searching of journals including CHI, FAccT, CSCW, *Journal on Responsible Computing*, *ACL*, *Big Data & Society*, and *AI & Society* contributed 21 sources (6%). Iterative gap-filling searches addressed underrepresented regions, concepts, or pipeline stages as the corpus took shape, contributing 31 sources (9%).

Inclusion criteria followed from the A/D/C scoping approach. Sources entered the corpus when they engaged community impacts substantively. Most sources additionally provided direct evidence about data production practices or AI systems. This meant including sources that analyzed AI system behavior, deployment, or evaluation in relation to community outcomes; examined data sourcing, processing, annotation, governance, or infrastructure with implications for affected communities; provided community-governed protocols, sovereignty statements, or governance frameworks; or established theoretical or epistemological foundations addressing power, resistance, extraction, or marginalization in ways essential for interpreting AI data practices and their consequences. We excluded sources that, for example, discussed AI ethics, fairness, or responsible AI at a high level without addressing data practices or community impacts; focused solely on model performance, technical optimization, or algorithmic advances without sociotechnical analysis; reported community-based research unrelated to AI systems or data production; or were non-English (a pragmatic limitation we discuss more below). Our inclusion criteria did not require sources to adopt a critical stance; sources that documented data production practices without evaluating their impacts also entered the corpus when they engaged substantively with community contexts. However, the C boundary condition—requiring engagement with community impacts—means the corpus foregrounds scholarship and practice that attends to affected populations, which favors justice-oriented work.

Quality assessment varied by source type and followed guidance from Kamei et al. [69]. Academic items underwent venue peer review. Grey-literature items required additional evaluation; we assessed organizational authority, author expertise, community recognition, and the provenance of policy documents, and we interpreted community outputs through alignment with decolonizing methodologies and community endorsement. These criteria track with grey-literature appraisal in multivocal reviews and draw on elements of Garousi et al. [49]'s framework, including stated aims, methodological clarity, contribution, and outlet type, which for this work primarily meant community-governed and sovereignty-oriented materials.

Screening proceeded in two stages: titles and abstracts were reviewed for relevance to the C boundary, followed by full-text assessment. The first author led database and grey literature searches. Two authors independently read full-text articles, prepared summaries, and presented sources in batches to the full team for consensus review. Disagreements on inclusion criteria and relevance to the A/D/C scoping approach were resolved through discussion. This process occurred in two rounds (August–September 2024), with each round reviewing approximately 100 candidate sources. These consensus rounds established shared standards before the two authors completed full screening of the 350-source corpus in February 2025. The final corpus contains 258 academic items (74%) and 92 grey-literature items (26%).

3.2 Corpus Creation

We created a datasheet that categorizes each source across multiple dimensions to provide maximum contextualization [34]. Coded categories included bibliographic metadata (author, year, venue, type), A/D/C coverage, pipeline stage, historical era, orientation, geographic focus, and author affiliation. For each source, we additionally recorded a unique rationale for A/D/C coverage and a brief summary to support traceability of sourcing and selection decisions. The same two authors who led source selection conducted categorization using the collaborative consensus approach established during screening.

Scoping Approach We categorized each source based on where direct evidence appears across our three-part inquiry: AI systems (A), data production practices (D), and community impacts (C). Every source engages all three dimensions analytically, but sources vary in what they directly support versus what requires interpretive connection. We wrote rationales stating what each source contributes to A, D, and C to clarify where direct evidence appears and where relevance is interpretive and to provide transparent disclosure of our interpretive stance on each source. Tags indicate where direct evidence is present. We do not tag A or D dimensions alone because all sources must engage community impacts (C) to enter the corpus. Our coding produced four categories, described in Table 2.

Pipeline Stages. We mapped each source to stages of a simplified AI development pipeline (Figure 1): *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) [93]. We mapped each source to a pipeline stage and sub-stage to make visible where specific mechanisms arise and how decisions at those points propagate through later phases—what prior work characterizes as cascading effects that compound downstream harms [128, 147]. Sources spanning multiple stages or describing cross-cutting dynamics were tagged accordingly.

Table 2. Coding definitions showing how corpus sources engage AI systems (A), data production practices (D), and community impacts (C)

Code	Description	Example Sources
ADC	Direct evidence relevant to AI systems, data production, and community impacts	Garcia et al. [48] on critical refusal as an intervention into extractive data logics and governance; Hall et al. [54] on participatory, community-engaged dataset production; Park et al. [111] on designing accessible infrastructures for collecting AI data from people with disabilities; Rifat et al. [126] on categorization politics and context erasure in annotating faith-based violence data; Lewis et al. [84] on Indigenous protocol–aligned dataset construction and culturally grounded AI applications
DC	Direct evidence for data production and community impacts; AI relevance is interpretive	Adley et al. [3] on ethical data collection with marginalized groups and power dynamics in practice; Cooper et al. [31] on community-collaborative research models emphasizing shared control and benefit; Hancock et al. [55] on tensions in data sharing and harms within a modern slavery data ecosystem; Taylor and Kukutai [151] on Indigenous Data Sovereignty and metadata governance; Pool [121] on colonial census practices replacing Māori knowledge systems
AD	Direct evidence for AI systems and data production; community impacts are clearly implied	Bhardwaj et al. [10] on evaluating ML datasets through a data-curation lens and FAIR principles; Koch et al. [77] on dataset reuse and benchmark concentration; Sambasivan et al. [128] on data cascades and hidden labor in high-stakes ML pipelines; Schiff et al. [135] on translating AI principles into practice via participatory, iterative impact assessment; Zhao et al. [170] on fairness-curation challenges faced by dataset curators across organizational and socio-political contexts
C	Direct evidence about community impacts only; A and D relevance is interpretive	Battiste [8] on Indigenous epistemologies and marginalization; Haraway [58] on situated knowledge and partial perspective; Igwe et al. [61] on non-extractive research principles; James [65] on colonial extraction economies and collective resistance in the Haitian Revolution; Shapiro and McNeish [140] on hyper-extractivism and resistance

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, though Era 1 practices appear retrospectively in multi-era sources (n=95, 27.1%) that trace historical continuities in data production. The concentration of sources in Era 3 (n=241, 68.9%) reflects the recency of both scholarly and practitioner attention to AI data production at industrial scale.

Orientations. Each source received a single orientation code reached through team consensus based on the primary analytical purpose the source served in this inquiry. *Extractive* sources provided direct evidence of practices undermining consent, compensation, or community benefit. *High-agency principles* advanced normative frameworks with explicit policy or governance recommendations. *High-agency practices* described operationalized initiatives with concrete implementation details.

Synthesis. We synthesized findings through iterative analysis across these dimensions. When multiple sources described similar mechanisms across different contexts, we consolidated these into recurring patterns. Individual

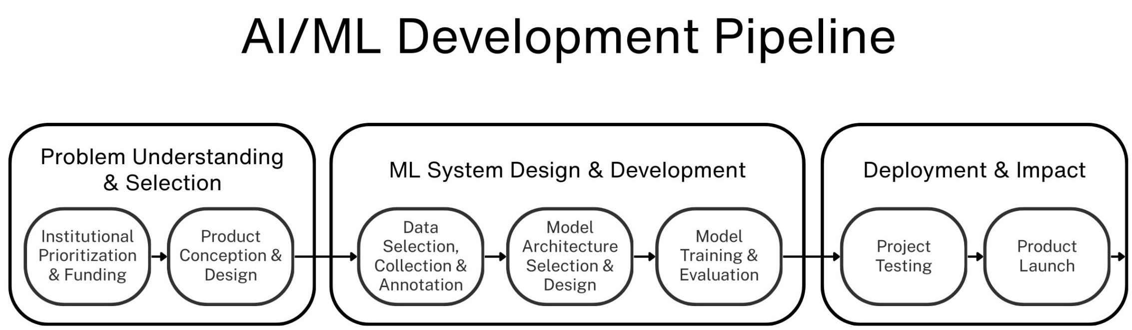

Fig. 1. Simplified AI development pipeline used to map corpus sources to the stages where specific mechanisms arise.

sources could exemplify multiple patterns. Patterns were distilled into the five analytic domains described in Section 4. Complete coding definitions with examples appear in the supplementary materials.

3.3 Limitations and Reflexivity

We recognize that subjectivity shapes our interpretations, though transparent documentation and multi-researcher validation helped address this inherent limitation. Connecting multiple disciplinary traditions, historical eras, and global contexts proved challenging, creating "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 Middle Eastern, 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. We acknowledge that English-language search restrictions inherently reinforce Western-centric representation. Scholarly and community infrastructures condition what becomes visible in review corpora, creating unevenness despite our efforts. Therefore, we believe it is important to consider our own identities alongside the analysis of this work given that our backgrounds and perspectives may bias the interpretation of this work [33].

The authors hold diverse racial and ethnic identities (Black, White, Mixed-race) with cultural roots in the United States, Canada, South Africa, Ghana, Japan, and France. These backgrounds shaped our ability to recognize and access specific community-led networks (particularly in African and Indigenous contexts) while leaving others less visible to us. In terms of epistemic lens, we work across industry and academia, with backgrounds in computer science, HCI, and the humanities. This dual positioning allowed us to bridge the gap between technical documentation and critical theory, for example, recognizing "grey literature" as rigorous evidence of high-agency alternatives. However, our location within these professionalized research institutions also means we likely missed grassroots resistance tactics that do not circulate in written or digital forms. We recognize that we are observing extraction from within the institutions that often facilitate it, and we present these findings as a necessary, though partial, mapping of the landscape.

4 Findings

We structure our findings according to five domains of data production, which we have conceptualized as analytic elements. Rather than logistical steps, these domains function as sites where power is negotiated, contested, and encoded: **Data Relations** (the negotiation of agency and terms of engagement), **Data Labor** (the creation versus capture of value), **Data Representation** (the exercise of epistemic authority through categorization), **Data Infrastructure** (the allocation of capacity and provenance), and **Data Governance** (the enforcement of sovereignty and accountability). Within each domain, we identify specific extractive mechanisms—technical or institutional habits that centralize control—and contrast them with high-agency pathways where communities are actively reclaiming authority.

Although mechanisms associated with each domain can appear at multiple points in AI development, consistent tendencies emerge across the corpus: decisions about relations often arise upstream as problems are framed; labor arrangements cluster within mid-stream annotation workflows; representational decisions crystallize where ontologies and preprocessing pipelines are defined; infrastructural conditions span stages but become most visible as systems scale; and governance concerns intensify downstream as models move toward evaluation and deployment. These tendencies help situate each domain without implying a fixed or linear pipeline.

Each domain is introduced through a small set of examples that surface the mechanisms we observed across the corpus. These examples are intended as points of entry into a broader landscape. The larger set of mappings, summaries, and domain categorizations is available in the datasheet for readers who wish to trace these patterns in greater depth.

4.1 Data Relations

Data relations define the structural terms of engagement between model developers and the communities from whom knowledge is derived. In mainstream industry discourse, these engagements are frequently reduced to legalistic questions of copyright compliance or static "terms of service." However, our corpus reveals that these legal frameworks often serve to obscure the underlying power dynamics [113]. Relations are not merely contractual; they are the primary site where agency is either stripped or substantiated. In extractive regimes, relations are characterized by the severance of ties between data and its creators; as Leanne Betasamosake Simpson articulates, "extraction removes all of the relationships that give whatever is being extracted meaning" [76]. High-agency relations, conversely, position data production as a negotiated partnership where community authority persists even after data is collected. Across eras, relational distance between data producers and affected communities has widened, from institutional mediation to platform-brokered terms of service to practices that bypass relational negotiation entirely.

The assumption of availability constitutes the primary mechanism of extractive relations. Technical workflows for foundation models frequently operate on the premise that any data accessible on the public web is a "standing reserve" available for ingestion. This logic converts public existence into implicit consent. Large-scale scraping initiatives, such as the corpora used to train models like CLIP [136, 137] or T5 [124], for example, bypass the negotiation of relationship entirely, including legally, by treating the act of publication as a forfeiture of rights [72, 133]. Relation-less forms of data production systematically ignore the contextual intent of the data creator, whether it be repurposing religious texts or intimate narratives as generic linguistic tokens, none of which are "just data" [59]. By removing the requirement to ask, the assumption of availability structurally precludes the possibility of refusal, rendering the relationship unilateral.

Transactional asymmetry reinforces this extraction by decoupling value generation from risk. This manifests in "digital extractivism," where Global Majority communities provide the raw material while the risks—such as the loss of privacy or the commodification of cultural heritage—are externalized back to them [64]. The dynamic functions through "accumulation by dispossession," where the terms of engagement are dictated by the extractor, treating communities as resources rather than partners [155]. Relational asymmetries are both economic and epistemic. AI developers gain a model of the world, while communities lose control over how they are represented within it, often leading to "opportunity loss" where resources are withheld based on extractive profiling [141].

High-agency relations counter these mechanisms by shifting from static terms of service to dynamic and revocable consent. Rather than viewing consent as a one-time gatekeeping mechanism, high-agency approaches frame it as an ongoing relationship. The Speech Accessibility Project [1] and other initiatives that engage disability communities aptly demonstrate how relationships can precede collection: communities are partners who co-define the terms of engagement before recruiting paid volunteers and help ensure the protocol aligns with community safety needs [1, 111]. Similarly, feminist frameworks for "embodied consent" argue for agreements that are specific, enthusiastic, and revocable, challenging the broad permissions usually buried in click-through agreements [146].

In Indigenous contexts, high-agency relations manifest as relational sovereignty. Te Hiku Media's approach to Māori data rejects the concept of open-source availability in favor of whanaungatanga (connection/relationship), where data access is determined by the strength and trust of the relationship between the parties [26, 57]. This reintroduces friction into the data pipeline by design: access is not a default state but a negotiated privilege that requires maintaining a relationship with the originating community [78]. By replacing the assumption of availability with permissioned access, these models force a structural acknowledgment of community agency.

Key takeaway: Data relations determine the flow of agency. Extractive mechanisms rely on the assumption of availability, treating public data as a resource to be mined and severing the link between creators and their data. This creates transactional asymmetry, where developers capture value while communities bear the risk. High-agency relations replace this with dynamic consent and relational sovereignty, ensuring that data production remains a negotiated partnership where community authority persists throughout the technical lifecycle.

4.2 Data Labor

Data labor encompasses the human energy and interpretive judgment required to bridge the gap between raw information and computational capability. While often obscured by the metaphor of "autonomous" AI, our corpus confirms that model performance remains strictly dependent on human workers who select, annotate, validate, and moderate content [24, 51, 83, 102]. In extractive regimes, this labor is characterized by value capture, where the semantic value generated by human judgment is stripped from the worker and concentrated in the model, often leaving the contributor with little to no recognition or economic return. High-agency approaches, conversely, frame labor as expertise, positioning annotators as skilled contributors whose situated knowledge is essential to system quality. The character of data labor has shifted across eras, from specialized domain expertise to distributed crowdwork to industrial-scale annotation and evaluation, with each transition further distancing the worker from the system's eventual use.

Invisibilization by design constitutes the primary mechanism of labor extraction. Dataset and platform architectures are frequently designed to present corpora as neutral technical artifacts rather than products of human judgment, masking the interpretive decisions embedded in every labeled example [94]. This structural opacity serves to commodify the worker; by decomposing complex cultural tasks into fragmented "microtasks," platforms strip the work of its context, rendering the worker interchangeable and the labor invisible [38]. Here, a structural design choice renders the human contribution indistinguishable from the system's output, with the upshot of systematically preventing workers from asserting authorship claims or contesting the terms of their participation.

Reciprocity failure reinforces this dynamic by extracting labor without returning value. This manifests most clearly in "unwitting" labor, where user interactions (e.g., solving CAPTCHAs, tagging photos, or correcting autocomplete suggestions) are harvested to train models without the user's explicit knowledge or compensation [15, 100]. In Global Majority contexts, this mechanism appears in the outsourcing of trauma-inducing content moderation or complex annotation to workers in low-income regions, who perform essential semantic labor for wages that do not reflect the cognitive intensity of the work [158]. The system is optimized to externalize the costs of dataset construction to the worker while centralizing the economic benefits.

High-agency labor counters these mechanisms by restructuring the economic and attributional relationship between modelers and workers. These approaches restore context and visibility to the labor process. The organization Karya, for example, demonstrates how data collection can function as a tool for economic redistribution; by establishing ethical wage floors and data ownership structures for rural Indian workers, they reframe annotation as a skilled, compensated profession [2]. Similarly, the Masakhane community creates participatory research models where African language speakers function not as passive data subjects, but as credited authors and technical collaborators throughout the pipeline [106]. Emerging initiatives like Ubuntu-AI attempt to encode these rights directly into the data lifecycle through profit-sharing mechanisms, ensuring that artists and creators retain a stake in the value their data generates [105].

Key takeaway: Data labor is economic and political. Extractive mechanisms rely on invisibilization by design, decomposing expert judgment into fragmented tasks to obscure the worker and facilitate value capture. This severs the link between labor and downstream value. High-agency approaches replace this with labor as expertise, ensuring that contributions are visible, attributed, and compensated as skilled work that persists within the technical system.

4.3 Data Representation

Data representation determines how communities become computationally visible. Representation is as much a question of inclusion ratios or diversity statistics as it is one of epistemic authority: who gets to define the categories, taxonomies, and labels that structure the digital world. In extractive regimes, representation creates visibility without power, often flattening complex, relational identities into rigid categories that facilitate control or consumption. High-agency approaches, conversely, frame representation as plural epistemologies, ensuring that data structures reflect community worldviews rather than forcing local knowledge into universalizing boxes. Representational dynamics have scaled across eras, from deliberate institutional categorization to crowdsourced labeling within inherited ontologies to web-scale ingestion that absorbs existing representational patterns without deliberate curation.

Ontological imposition constitutes the primary mechanism of representational extraction. Institutional problem formulation often imposes external taxonomies on communities before they even enter the pipeline. This manifests as "data universalism" [97] where Western logics of property and individualism are treated as neutral defaults, overwriting Indigenous ontologies that emphasize relationality and collective stewardship [84]. For example, psychological frameworks developed in "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic) contexts fail to map onto collective ontologies, yet are deployed globally as standard [40, 96, 99]. Consequently, even when diverse data is collected it is structurally distorted to fit the model's worldview, rendering specific cultural meanings "absent" even within inclusion efforts [11].

Context stripping reinforces this dynamic during annotation and processing. To make data "model-ready," complex human experiences must be converted into discrete labels. This process often relies on "lazy" data practices that collapse distinct protected attributes like race and ethnicity into coarse categories to satisfy technical constraints, erasing intersectional realities [142]. Annotation workflows that lack community-defined criteria force workers to resolve ambiguity by falling back on institutional defaults, which appear neutral but encode specific cultural biases [132].

Automated filtering pipelines compound this by removing content that signals non-normative identities under the guise of "cleaning," disproportionately purging data from non-Western contexts or disability communities [91].

Synthetic displacement introduces a new mechanism of extraction: representation without presence. As privacy regulations tighten, developers increasingly turn to synthetic data (e.g., fabricated medical records, artificial faces, and simulated identities) to populate datasets. While this bypasses the need for individual consent [82], it severs the link between representation and reality. Communities become represented in systems they never participated in, inheriting the risks of misidentification or caricature without any pathway to contest how they are depicted [163]. The resulting "diversity-washing" effect is such that models appear inclusive while structurally excluding actual community members.

High-agency representation counters these mechanisms by building pluralistic and community-grounded corpora. These initiatives prioritize depth and context over scale. For instance, the Abundant Intelligences project reimagines AI development through Indigenous knowledge systems, refusing to separate data from the land and relations that generate it [85]. Similarly, examples from Africa and Oceania demonstrate how regional collaborations can curate datasets that serve local linguistic needs—such as the InkubaLM model—rather than adapting to global benchmarks [43, 157]. By maintaining representational authority, these projects ensure that visibility serves community goals, such as language revitalization, rather than external commodification.

Key takeaway: Data representation is epistemic and political. Extractive mechanisms rely on ontological imposition and context stripping, imposing external taxonomies and flattening meaning to fit technical defaults. This treats visibility as neutral even when it creates exposure. High-agency approaches replace this with plural epistemologies, grounding representation in community-defined categories and preserving the specificity of local knowledge against universalizing standards.

4.4 Data Infrastructure

Data infrastructure allocates capacity and determines where data lives, who controls access, and how material circulates across model pipelines. While often treated as neutral plumbing designed for efficiency, infrastructure emerges in the literature as a primary site of political contestation. In extractive regimes, infrastructure is configured to maximize velocity and volume, creating technical conditions where consent and context are structurally impossible to maintain. High-agency approaches, conversely, design for traceability and distribution, ensuring that community authority travels with the data. Infrastructure has consolidated across eras, from locally held institutional datasets to cloud-hosted repositories to globally indexed architectures that concentrate access while diffusing accountability.

Centralization without governance configures extraction at an industrial scale. Foundation-model development relies on automated pipelines that ingest content from large-scale web sources such as Common Crawl or LAION to maximize throughput [42, 136, 137]. This configuration privileges actors with substantial compute resources and treats data availability as a default condition. The asymmetry is infrastructural: collection mechanisms operate at speeds that make oversight and contestation structurally unworkable for data subjects [164].

Benchmark infrastructures act as gatekeeping mechanisms that enforce dominant (Western) epistemologies as universal standards. Reliance on a narrow set of legacy datasets, such as ImageNet [37] and MS COCO [88], entrenches specific linguistic, cultural, and demographic assumptions as infrastructural norms [38, 77]. Because creating culturally specific alternatives requires substantial institutional support, Euro-American category systems persist as de facto standards through infrastructural path dependence [77].

Provenance compression serves as the third mechanism, severing datasets from their originating communities and the relational contexts of their creation. Contemporary web-scrape datasets often operate through severe documentation gaps—reinforcing "web-as-platform" assumptions that treat public accessibility as permission to extract [133]. Infrastructure that treats provenance as optional enables downstream actors to shift responsibility for data quality and rights onto untraceable contributors [89].

High-agency infrastructure counters these mechanisms by embedding community-defined constraints directly into technical architectures. Federated and distributed systems shift authority by enabling collaboration without centralizing data. Emerging frameworks for "data spaces" allow communities to retain local control over storage and access while supporting model development [47, 60]. Similarly, stewardship-based architectures like Masakhane's distributed research platforms operationalize co-designed metadata standards, ensuring that data does not become "loose" but remains tethered to its community of origin [106].

Key takeaway: Data infrastructure is about capacity and provenance. Extractive architectures rely on centralization without governance to maximize velocity and provenance compression to sever data from its originating obligations. This makes extraction structurally easy and accountability expensive. High-agency alternatives deploy federated and distributed systems, redistributing capacity so that community authority remains technically enforceable as data circulates.

4.5 Data Governance

Governance establishes the rule-sets that authorize data production: it determines when collection is legitimate, what contextual grounding is required, and who holds authority over circulation. These rules operate upstream of participation, labor, and representation, guiding the conditions under which data production becomes legitimate. High-asymmetry governance frameworks create wide discretionary space for extractive practices, whereas high-agency governance embeds community control directly into the structures that shape data lifecycles. Governance challenges have intensified across eras as data flows outpaced regulatory frameworks, from institutional norms governing small datasets to platform policies to global-scale extraction operating across jurisdictional boundaries.

Regulatory arbitrage constitutes the primary mechanism of extractive governance. Often termed "ethics dumping," this practice exploits fragmented global regulations to harvest data in regions with weaker protections, converting behavioral interactions into institutional assets without oversight [153]. This dynamic transforms regulatory variation into a resource for extraction: vulnerable populations in low- and middle-income countries may receive limited digital services (like Facebook's Free Basics) in exchange for extensive, uncompensated data harvesting [107]. Coercive collection in humanitarian settings, such as biometric registration in Ethiopian refugee camps, further illustrates how governance gaps allow institutions to bypass the consent standards required in their home jurisdictions [154].

Open-loop extraction reinforces this asymmetry by decoupling deployment from accountability. Models trained on narrow, Western-centric data are frequently deployed globally, shifting the burden of performance failures—such as diagnostic errors in healthcare AI—onto underserved communities [6, 108]. This mechanism externalizes risk: communities excluded from the governance of training data nonetheless become sources of performance feedback during deployment. Their interactions refine the system, yet they possess no authority to challenge the model's adequacy or recall the data they generate [156]. Governance here functions to protect the model developer's intellectual property while leaving the data subject's sovereignty unprotected.

High-agency governance counters these mechanisms through sovereignty-based licensing and critical refusal. Rather than relying on open-access defaults, these approaches encode community authority into the legal terms of the data itself. The Kaitiakitanga License, developed by Te Hiku Media, exemplifies this by legally binding data usage to Māori tikanga (protocols), preventing extractive reuse by third parties [152]. Similarly, the Esethu Framework for African language data establishes sovereignty provisions that mandate community benefit-sharing and protect annotators [125]. Beyond licensing, critical refusal operates as a form of affirmative governance. By setting ex-ante boundaries on participation, communities assert that unreadability is a safety condition. Longstanding tactics of opacity and masking establish practical limits on what institutions may extract [19]. When viewed as governance, refusal is not a lack of data; it is an enforcement of sovereignty that limits extractive reach by design [48].

Key takeaway: Data governance distinguishes accountability from exploitable discretion. Extractive mechanisms rely on regulatory arbitrage (ethics dumping) and open-loop extraction, engaging in collection without contextual grounding and turning deployment into unconsented data acquisition. High-agency approaches establish sovereignty-based licensing and critical refusal, creating enforceable preconditions that align data production with community-defined control and embed agency beyond the point of collection.

5 Discussion

Our analysis of 350 sources across academic and grey literature reveals that AI data production is not merely a logistical preliminary to model development, but a distinct sociotechnical site where power is negotiated, contested, and encoded. By synthesizing evidence across AI systems, data production practices, and community impacts, we identify a clear divergence: extractive practices that prioritize scale, opacity, and labor externalization, versus high-agency pathways that prioritize relationality, sovereignty, and context. Notably, the five analytic domains we identify do not distribute evenly across the ML pipeline. Instead, the sources that comprise each domain cluster around the structural moments where key mechanisms take effect: **Data Relations** concentrates upstream in problem formulation and data selection; **Data Labor** anchors mid-pipeline annotation and enrichment; **Data Representation** spans early- to mid-pipeline ontology and preprocessing; **Data Infrastructure** forms a cross-cutting substrate most visible in mid-to-downstream development; and **Data Governance** clusters downstream where deployment, accountability, and sovereignty become salient. This patterned distribution indicates that extractive dynamics are not random but structurally embedded within distinct, yet interrelated, pipeline junctures.

These findings carry implications across responsible computing. Within HCI specifically, researchers have successfully interrogated downstream AI interaction and mid-stream model behavior, yet the specific mechanisms connecting upstream data production to downstream community impacts remain underexamined across computing research. Below, we discuss how researchers and practitioners can operationalize high-agency practices by treating data production as a primary site of design intervention. In doing so, we build on and extent scholarship in HCI and adjacent fields that examines data production through the lens of data laborers and data subjects [70, 71, 95, 131, 134, 159].

5.1 Reframing Data Production as "Upstream" Design

Our findings challenge the widespread norm—held by industry practitioners and, often implicitly, by their critics—of treating data as "found" infrastructure (Era 3). Instead, the evidence suggests data production is a series of design decisions—regarding relations, labor, and representation—that are often irreversible once encoded into a model. This recognition prompts us to argue that the "user" in human-centered AI must expand to include the data contributor—the artist, the annotator, the community member—whose agency is often circumvented by upstream infrastructure. Within HCI, this circumvention has predominantly been investigated in studies of data labor, which reveal how data annotators are frequently reduced to an interchangeable resource thereby constraining their subjectivities and interpretive work [70, 95, 159]. Building on this work, our review makes clear the various design choices like crowdsourcing interfaces that atomize tasks to obscure the worker's context (Data Labor) or scraping pipelines that strip provenance metadata (Data Infrastructure) which can circumvent agency and enforce extraction by design.

For HCI, this implies that data curation is a form of interaction design. The high-agency pathways our review surfaces make clear that alternative designs are possible. The community-led initiatives such as Masakhane's participatory NLP [92] or Māori data sovereignty protocols [78, 103] succeed not by "fixing" extraction after the fact, but by designing relational friction into the process. They replace the seamless, frictionless extraction of web scraping with protocols that require consent, negotiation, and maintenance.

5.2 Toward High-Agency Practices: Implications for HCI

Moving beyond critique, our analysis of high-agency pathways points toward concrete mechanisms for less-extractive AI development. We map these implications to three key shifts for research and practice.

5.2.1 From Universal Representation to Pluralistic "Small Data". The dominance of massive, web-scraped corpora (Era 3) enforces a universalizing worldview that erases minority contexts and agency in general. Our findings suggest that "de-biasing" these massive datasets is often less effective than building smaller, community-sovereign corpora. Indeed, critiques of contemporary efforts to build more "inclusive" or "de-biased" technologies often highlight a technosolutionist trap whereby the issue is purportedly addressed through large-scale capture of data about a community or culture without meaningful agency [23, 122, 123]. Our review reinforces how failing to allow communities to set the terms of inclusion for their data can inadvertently perpetuate extraction under the guise of inclusion. This extends critiques of "fair" AI that don't fundamentally shift power [68]. In contrast, high-agency pathways highlighted in our review demonstrate how different actors and groups are proactively responding to extractive data capture by imagining and building alternatives. For instance, efforts by Te Hiku Media to develop community led language datasets and technologies [43, 152] and the Community-driven African Next Voices project [169] directly counter efforts by big tech to seek out and capture cultural knowledge and data, by instead keeping data governed by communities. While authors, organizations, and initiatives offer unique contributions, their coordination and totality points to systemic alternatives that start with community needs and maintain community control, thus creating their own conditions for thriving rather than adapting to external constraints. By developing 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 research community has an opportunity to advance high-agency efforts by investing in federated data spaces rather than centralized lakes. We need infrastructure that allows models to learn from community data without that data ever leaving the community's local storage or jurisdiction (e.g., federated learning tailored for Indigenous sovereignty [47]). Furthermore, valuation metrics in AI research must shift away from scale at the expense of care [56, 66, 131, 168]. We encourage the HCI community to value (and publish) contributions that curate high-context, small-scale datasets with clear governance protocols, rather than rejecting them for lacking the scale of foundation model benchmarks.

5.2.2 From Transactional Labor to Relational Provenance. Our review highlights reciprocity failure and labor invisibility as central extractive patterns. This transactional model commodifies the work of data laborers—including annotators, content creators, and community members—distancing those performing the work from those capturing the value and contributing to the perceived "magic" of AI [129]. This is further complicated by opaque data collection practices that often make the data creator unaware of their contribution. While there is growing interest in data provenance as a key intervention point for mitigating harm of AI technologies [89, 162], our review affirms how the current dominant paradigm of data production is in tension with this end goal. We argue that this tension is, in part, a design challenge and call upon the research community to explore how data capture and sharing platforms might implement provenance-tracking mechanisms by design.

A provenance-first design approach could involve binding labor and authorship metadata to individual data points so that creators retain "credits" (similar to the Ubuntu-AI model [105]) that persist through the pipeline. More broadly, there is growing recognition that data annotation is fundamentally subjective and interpretive, often shaped by the sociocultural backgrounds and lived experiences of annotators [39, 138]. This motivates the design of data annotation processes and infrastructure that allows workers to signal ambiguity, refuse tasks that violate community norms, and capture disagreements in a structured form rather than forcing a choice that flattens cultural context. By doing so, researchers can enable downstream pluralistic modeling approaches that can handle meaningful divergences in perspectives [101].

5.2.3 From Open-Loop Extraction to Closed-Loop Governance. The governance gaps identified in our review show that once data is scraped, communities often lose control. In contrast, community-led governance approaches exemplify an alternative. For example, Māori data sovereignty frameworks in Aotearoa New Zealand demonstrate a coordinated ecosystem: 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. Such agency-oriented practices require dynamic consent and enforceable boundaries and necessitate technical implementations of Sovereignty-Based Licensing. Researchers can develop and standardize machine-readable licenses (similar to Creative Commons but for ML training) that explicitly forbid certain downstream uses (e.g., military application, generative mimicry) and trigger benefit-sharing clauses [125, 152].

To operationalize this, Institutional Review Boards (IRBs) and conference ethics reviews must look upstream, enforcing data transparency standards that treat data collection as a distinct object of ethical inquiry [9, 50]. Within academic peer review, data production is increasingly within scope of ethical inquiry. However, the focus remains largely on individual privacy and consent within papers presenting novel datasets, rather than deeper inquiries into the conditions under which data is produced and the extent to which communities retain any rights of refusal. This necessitates new review paradigms that prioritizes community consent in addition to individual terms of service, echoing calls for power-aware approaches that allows communities to attest and refuse data extraction [48, 68]. By scrutinizing these data cascades at the source [128], the review process can identify where the agency of the data contributor has been circumvented.

5.3 Methodological Contributions: The Value of Multivocality

Finally, this paper illustrates the utility of the Multivocal Literature Review (MLR) for investigating sociotechnical harm. The corpus shows systematic differences in what each source type contributes. Grey literature documents high-agency practices at 37.0% compared to 22.9% in white literature and draws from more diverse author affiliations, including NGO and non-profit organizations (17.4% vs. 1.6%). Grey sources also provide greater geographic specificity: only

26.1% lack a regional focus, compared to 48.8% of white literature sources, with particularly strong representation of African (17.4% vs. 8.5%) and Asia-Pacific contexts (8.7% vs. 2.7%). These patterns indicate that grey literature surfaces practitioner knowledge and regionally grounded perspectives underrepresented in formal publication channels. In short, a significant portion of our high-agency evidence came not from peer-reviewed academic venues, but from "grey" literature—community manifestos, tribal resolutions, and worker inquiries. Limiting the scope to academic literature alone would likely have surfaced the harms of extraction, which are well-documented in academia, while underrepresenting the alternatives documented in policy and community organizing. For responsible computing research, this underscores that "state-of-the-art" knowledge regarding justice and equity often resides outside the academy, as does the general "state-of-practice." Future work on AI harms would benefit from adopting multivocal methods to ensure that community-generated resistance and innovation are recognized as rigorous evidence.

6 Conclusions

As AI development consolidates around foundation models trained on internet-scale scrapes, the risk of deepening extractive relations is acute. However, this trajectory is not inevitable. By analyzing the data production pipeline through the lens of Data Relations, Data Labor, Data Representation, Data Infrastructure, and Data Governance, we see that every dataset is a record of power relations. This paper contributes a mapping of these relations, offering researchers and practitioners a diagnostic tool to identify extraction and a catalog of precedents for resistance. The shift to less-extractive AI requires more than better algorithms; it requires designing the upstream sociotechnical infrastructures that determine whose knowledge counts, how it is valued, and who governs its future. Our review affirms that a less-extractive future is not merely an aspiration; it is actively being built by communities pursuing alternatives to the status quo.

References

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

[2] 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/

[3] Mark Adley, Hayley Alderson, Katherine Jackson, William McGovern, Liam Spencer, Michelle Addison, and Amy O'Donnell. 2024. Ethical and practical considerations for including marginalised groups in quantitative survey research. *International Journal of Social Research Methodology* 27, 5 (2024), 559–574. doi:10.1080/13645579.2023.2228600

[4] Leah Hope Ajmani, Jasmine C. Foriest, Jordan Taylor, Kyle Pittman, Sarah Gilbert, and Michael Ann DeVito. 2024. Whose Knowledge is Valued? Epistemic Injustice in CSCW Applications. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 523:1–523:28. doi:10.1145/3687062

[5] 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

[6] 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].

[7] Stefan Baack. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2199–2208. doi:10.1145/3630106.3659033

[8] 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

[9] 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

[10] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine learning data practices through a data curation lens: An evaluation framework. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1055–1067. doi:10.1145/3630106.3658955

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

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

[13] 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

[14] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 948–958. doi:10.1145/3531146.3533157

[15] 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

[16] 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

[17] John R. Bowman. 1989. *Capitalist Collective Action: Competition, Cooperation and Conflict in the Coal Industry.* Cambridge University Press. Google-Books-ID: fnl6sAYpRLYC.

[18] 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

[19] Simone Browne. 2015. *Dark Matters: On the Surveillance of Blackness.* Duke University Press. doi:10.1215/9780822375302

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[21] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512. doi:10.1177/2053951715622512

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

[23] Alan Chan, Chinasa T Okolo, Zachary Terner, and Angelina Wang. 2021. The limits of global inclusion in AI development. *arXiv preprint arXiv:2102.01265* (2021).

[24] Srravya Chandhiramowuli, Alex S. Taylor, Sara Heitlinger, and Ding Wang. 2024. Making Data Work Count. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 90:1–90:26. doi:10.1145/3637367

[25] 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

[26] Donavyn Coffey. 2021. Māori are trying to save their language from Big Tech. *Wired* (April 2021). https://www.wired.com/story/maori-language-tech/

[27] 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.

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

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

[30] Common Crawl. 2025. *Common Crawl*. https://commoncrawl.org/

[31] 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

[32] Matthew Cotton. 2017. Fair fracking? Ethics and environmental justice in United Kingdom shale gas policy and planning. *Local Environment* 22, 2 (Feb. 2017), 185–202. doi:10.1080/13549839.2016.1186613

[33] Payton Croskey, Fabian Offert, Jennifer Jacobs, and Kai M. Thaler. 2025. Liberatory Collections and Ethical AI: Reimagining AI Development from Black Community Archives and Datasets. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 900–913. doi:10.1145/3715275.3732058

[34] 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

[35] 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].

[36] 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.

[37] 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

[38] 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

[39] Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. arXiv:2112.04554 (Dec. 2021). doi:10.48550/arXiv.2112.04554 arXiv:2112.04554 [cs].

[40] 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

[41] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (Jan. 2021). https://par.nsf.gov/biblio/10341458-retiring-adult-new-datasets-fair-machine-learning

[42] 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].

[43] 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

[44] Pedro Ferreira. 2024. Examining the "Local" in ICT4D: A Postcolonial Perspective on Participation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3613904.3642748

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

[46] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (2024), 1097–1179. doi:10.1162/coli_a_00524

[47] 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).

[48] Patricia Garcia, Tonia Sutherland, Niloufar Salehi, Marika Cifor, and Anubha Singh. 2022. No! Re-imagining Data Practices Through the Lens of Critical Refusal. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 315:1–315:20. doi:10.1145/3557997

[49] 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

[50] 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

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

[52] Michael Haenlein and Andreas Kaplan. 2019. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review* 61, 4 (Aug. 2019), 5–14. doi:10.1177/0008125619864925

[53] 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

[54] 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].

[55] Jamie Hancock, Sarada Mahesh, Jennifer Cobbe, Jatinder Singh, and Anjali Mazumder. 2024. The tensions of data sharing for human rights: A modern slavery case study. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 974–987. doi:10.1145/3630106.3658949

[56] Alex Hanna and Tina M. Park. 2020. Against Scale: Provocations and Resistances to Scale Thinking. arXiv:2010.08850 (Nov. 2020). doi:10.48550/arXiv.2010.08850 arXiv:2010.08850 [cs].

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

[58] 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

[59] 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].

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

[61] Paul Agu Igwe, Nnamdi O. Madichie, and David Gamariel Rugara. 2022. Decolonising research approaches towards non-extractive research. *Qualitative Market Research: An International Journal* 25, 4 (Jan. 2022), 453–468. doi:10.1108/QMR-11-2021-0135

[62] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 1311–1320. doi:10.1145/1753326.1753522

[63] Victor Chidubem Iwuoha. 2025. European Biometric Borders and (Im)Mobilities in West Africa: Reflections on Migrant Strategies for Border Circumvention and Subversion. *Politics Policy* 53, 1 (2025), e12653. doi:10.1111/polp.12653

[64] 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/

[65] 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.

[66] 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

[67] James H Jones. 2008. The Tuskegee syphilis experiment. *The Oxford textbook of clinical research ethics* (2008), 86–96.

[68] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169–169.

[69] Fernando Kamei, Igor Wiese, Gustavo Pinto, Waldemar Ferreira, Márcio Ribeiro, Renata Souza, and Sérgio Soares. 2022. Assessing the Credibility of Grey Literature: A Study with Brazilian Software Engineering Researchers. *Journal of Software Engineering Research and Development* 10 (june 2022). doi:10.5753/jserd.2022.1897

[70] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. doi:10.1145/3544548.3580645

[71] Reishiro Kawakami and Sukrit Venkatagiri. 2024. The Impact of Generative AI on Artists. In *Proceedings of the 16th Conference on Creativity & Cognition (Camp;C '24)*. Association for Computing Machinery, New York, NY, USA, 79–82. doi:10.1145/3635636.3664263

[72] 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

[73] 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

[74] 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

[75] 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

[76] 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

[77] 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].

[78] 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

[79] 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].

[80] 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

[81] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (Nov. 1998), 2278–2324. doi:10.1109/5.726791

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

[83] Tuukka Lehtiniemi and Minna Ruckenstein. 2022. *Prisoners training AI: Ghosts, Humans and Values in Data Labour.* Routledge, Abingdon, 184–196. doi:10.4324/9781003170884-16

[84] 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 Kealiikanakaoleo-haililani, 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

[85] 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

[86] 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

[87] 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

[88] 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.

[89] 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].

[90] Eneko Lopez, Jaione Etxebarria-Elezgarai, Jose Manuel Amigo, and Andreas Seifert. 2023. The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples. *Analytica Chimica Acta* 1275 (2023), 341532. doi:10.1016/j.aca.2023.341532

[91] 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

[92] 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

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

[94] 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

[95] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–25. doi:10.1145/3415186

[96] Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why AI is WEIRD and shouldn't be this way: towards AI for everyone, with everyone, by everyone. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25, Vol. 39).* AAAI Press, 28657–28670. doi:10.1609/aaai.v39i27.35092

[97] 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

[98] 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

[99] 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

[100] 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

[101] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110. doi:10.1162/tacl_a_00449

[102] James Muldoon, Callum Cant, Boxi Wu, and Mark Graham. 2024. A Typology of AI Data Work. *Big Data and Society* 11, 11 (March 2024). doi:10.1177/20539517241232632

[103] 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

[104] Diego I. Murguía and Kathrin Böhling. 2013. Sustainability reporting on large-scale mining conflicts: the case of Bajo de la Alumbrera, Argentina. *Journal of Cleaner Production* 41 (Feb. 2013), 202–209. doi:10.1016/j.jclepro.2012.10.012

[105] 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/

[106] 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

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

[108] 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

[109] 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

[110] Ciaran O'Faircheallaigh. 2015. Social Equity and Large Mining Projects: Voluntary Industry Initiatives, Public Regulation and Community Development Agreements. *Journal of Business Ethics* 132, 1 (Nov. 2015), 91–103. doi:10.1007/s10551-014-2308-3

[111] 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

[112] Yong Jin Park and S. Mo Jones-Jang. 2023. Surveillance, Security, and Ai as Technological Acceptance. *AI and Society* 38, 6 (2023), 2667–2678. doi:10.1007/s00146-021-01331-9

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

[114] Charlotte Paul and Barbara Brookes. 2015. The Rationalization of Unethical Research: Revisionist Accounts of the Tuskegee Syphilis Study and the New Zealand "Unfortunate Experiment". *American Journal of Public Health* 105, 10 (Oct. 2015), e12–19. doi:10.2105/AJPH.2015.302720

[115] 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

[116] Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts. https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts

[117] 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].

[118] 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

[119] Sidney Perkowitz. 2021. The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing* Winter 2021 (Feb. 2021). doi:10.21428/2c646de5.62272586

[120] Claudio Santos Pinhanez and Edem Wornyo. 2025. Ethical Co-Development of AI Applications with Indigenous Communities. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3706599.3706649

[121] Ian Pool. 2016. *Colonialism's and postcolonialism's fellow traveller: the collection, use and misuse of data on indigenous people* (1st ed.). ANU Press. doi:10.22459/CAEPR38.11.2016.04

[122] Rida Qadri, Michael Madaio, and Mary L. Gray. 2025. Confusing the Map for the Territory – Communications of the ACM. https://cacm.acm.org/opinion/confusing-the-map-for-the-territory/

[123] Rida Qadri, Piotr Mirowski, and Remi Denton. 2025. AI and Non-Western Art Worlds: Reimagining Critical AI Futures through Artistic Inquiry and Situated Dialogue. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–17. doi:10.1145/3706598.3714049

[124] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1 (Jan. 2020), 140:5485–140:5551.

[125] 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].

[126] 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

[127] 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

[128] 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

[129] Advait Sarkar. 2023. Enough With "Human-AI Collaboration". In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–8. doi:10.1145/3544549.3582735

[130] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. arXiv:2502.18682 (April 2025). doi:10.48550/arXiv.2502.18682 arXiv:2502.18682 [cs].

[131] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–37. doi:10.1145/3476058

[132] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 58:1–58:35. doi:10.1145/3392866

[133] 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

[134] Morgan Klaus Scheuerman, Allison Woodruff, and Jed R. Brubaker. 2025. How Data Workers Shape Datasets: The Role of Positionality in Data Collection and Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 9, 7 (Oct. 2025), CSCW300:1–CSCW300:42. doi:10.1145/3757481

[135] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. *Principles to Practices for Responsible AI: Closing the Gap.* doi:10.48550/arXiv.2006.04707 arXiv:2006.04707 [cs].

[136] 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

[137] 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

[138] Candice Schumann, Gbolahan O. Olanubi, Auriel Wright, Ellis Monk Jr., Courtney Heldreth, and Susanna Ricco. 2023. Consensus and Subjectivity of Skin Tone Annotation for ML Fairness. https://arxiv.org/abs/2305.09073v3

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

[140] Judith Shapiro and John-Andrew McNeish. 2021. *Our Extractive Age: Expressions of Violence and Resistance* (1 ed.). Routledge, London. doi:10.4324/9781003127611

[141] 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].

[142] 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

[143] 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).

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

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

[146] 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

[147] 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].

[148] 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.

[149] 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

[150] 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

[151] John Taylor and Tahu Kukutai (Eds.). 2016. *Indigenous Data Sovereignty: Toward an Agenda*. ANU Press, Acton, ACT, Australia. doi:10.22459/CAEPR38.11.2016

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

[153] 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

[154] 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/

[155] 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

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

[157] 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].

[158] 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].

[159] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation. arXiv:2203.10748 (March 2022). doi:10.48550/arXiv.2203.10748 arXiv:2203.10748 [cs].

[160] Lining Wang, Vaishnav Kameswaran, and Hernisa Kacorri. 2025. Toward a Taxonomy of Algorithmic Harms for Disability: A Systematic Review. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 3 (Oct. 2025), 2649–2665. doi:10.1609/aies.v8i3.36745

[161] Keren Weitzberg. 2025. Keeping people out of camps: biometric technologies, contested sovereignty, and border practices within humanitarian spaces. *Journal of Ethnic and Migration Studies* 51, 14 (Aug. 2025), 3590–3609. doi:10.1080/1369183X.2025.2513155

[162] Karl Werder, Balasubramaniam Ramesh, and Rongen (Sophia) Zhang. 2022. Establishing Data Provenance for Responsible Artificial Intelligence Systems. *ACM Transactions on Management Information Systems* 13, 2 (2022), 1–23. doi:10.1145/3503488

[163] 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

[164] David Gray Widder. 2024. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1295–1304. doi:10.1145/3630106.3658973

[165] 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

[166] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, London England United Kingdom, 1–10. doi:10.1145/2601248.2601268

[167] Lucy C. Woodall, Sheena Talma, Oliver Steeds, Paris Stefanoudis, Marie-May Jeremie-Muzungaile, and Alain de Comarmond. 2021. Co-development, co-production and co-dissemination of scientific research: a case study to demonstrate mutual benefits. *Biology Letters* 17, 4 (April 2021), 20200699. doi:10.1098/rsbl.2020.0699

[168] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday* (April 2024). doi:10.5210/fm.v29i4.13642

[169] 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

[170] Dora Zhao, Morgan Klaus Scheuerman, Pooja Chitre, Jerone T. A. Andrews, Georgia Panagiotidou, Shawn Walker, Kathleen H. Pine, and Alice Xiang. 2024. A taxonomy of challenges to curating fair datasets. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24, Vol. 37)*. Curran Associates Inc., Red Hook, NY, USA, 97826–97858.

Received 12 February 2026; revised 5 June 2009; accepted 5 June 2009