When Care Means Extraction: How Data Colonialism Is Legitimated in Humanitarian and Development Contexts

Journal:	*Big Data & Society*
Manuscript ID	Draft
Manuscript Type:	Original Research Article
Keywords:	technocolonialism, AI ethics, corporate data infrastructure, humanitarian technology, data colonialism, Technohumanism
Abstract:	Technology companies increasingly provide digital infrastructure to humanitarian and development organisations in the Global South, presenting these partnerships through vocabularies of protection, inclusion, efficiency, and connectivity. This article asks how that framing operates. Drawing on discourse analysis of approximately 170 publicly available documents across eight cases and four data practices (analytics, connectivity, cloud infrastructure, and digital identification), the article finds that the language used to justify corporate data collection does not simply describe a beneficial relationship. It produces the conditions under which extraction appears necessary and refusal irrational. The vocabularies attached to each data practice define affected communities as recipients of care rather than as agents who might contest the data relationship, and although different registers operate across practices, from "evidence-based protection" in predictive analytics to "legal identity for all" in digital identification, each produces the same result. The four data practices form a connected infrastructure in which each enables the next, routing data from communities in the Global South to Northern corporate servers while the language of care flows in the opposite direction. No data practice in the corpus has been dismantled after its original justification expired. The article advances data colonialism theory by demonstrating that extraction is sustained not only through structural conditions but through a discursive mechanism that renders extraction legible as care. The findings also challenge consent-based AI ethics frameworks, which presuppose a subject who can refuse, a condition this analysis shows does not hold when the alternative to data collection is exclusion from life-sustaining services.

When Care Means Extraction: How Data Colonialism Is Legitimated in Humanitarian and Development Contexts

Abstract

Technology companies increasingly provide digital infrastructure to humanitarian and development organisations in the Global South, presenting these partnerships through vocabularies of protection, inclusion, efficiency, and connectivity. This article asks how that framing operates. Drawing on discourse analysis of approximately 170 publicly available documents across eight cases and four data practices (analytics, connectivity, cloud infrastructure, and digital identification), the article finds that the language used to justify corporate data collection does not simply describe a beneficial relationship. It produces the conditions under which extraction appears necessary and refusal irrational. The vocabularies attached to each data practice define affected communities as recipients of care rather than as agents who might contest the data relationship, and although different registers operate across practices, from "evidence-based protection" in predictive analytics to "legal identity for all" in digital identification, each produces the same result. The four data practices form a connected infrastructure in which each enables the next, routing data from communities in the Global South to Northern corporate servers while the language of care flows in the opposite direction. No data practice in the corpus has been dismantled after its original justification expired. The article advances data colonialism theory by demonstrating that extraction is sustained not only through structural conditions but through a discursive mechanism that renders extraction legible as care. The findings also challenge consent-based AI ethics frameworks, which presuppose a subject who can refuse, a condition this analysis shows does not hold when the alternative to data collection is exclusion from life-sustaining services.

Keywords: *technocolonialism, AI ethics, corporate data infrastructure, Global South, dependency chain, humanitarian technology, technohumanitarianism*

Word count: 9,902

When Care Means Extraction: How Data Colonialism Is Legitimated in Humanitarian and Development Contexts

1. Introduction

In February 2019, the World Food Programme announced a partnership with Palantir Technologies valued at $45 million, granting the data analytics firm access to the personal information of over 90 million aid recipients across multiple crisis zones (Parker, 2019). The partnership was framed as a transformation of humanitarian delivery through data-driven efficiency. Yet the same company simultaneously provided analytical infrastructure to US Immigration and Customs Enforcement, an agency implicated in documented human rights violations against displaced communities (Amnesty International, 2020). The technology presented as beneficial in one context enabled enforcement against the same populations in another. The term "care" in this article does not invoke the ethics of care tradition or affective relationships between individuals. It refers to the institutional vocabulary through which corporate and humanitarian actors claim that a data practice serves the welfare of affected populations, presenting extraction as service. The critical humanitarianism literature has established that such claims are analytically significant, because care and governance are co-constitutive dimensions of humanitarian action (Barnett, 2011, p. 12). The question this article asks is how that co-constitution operates at the level of specific legitimating vocabularies.

Across the African continent, Facebook's Free Basics programme expanded into 32 countries, offering zero-rated internet access to populations with limited connectivity. The programme was presented as a philanthropic response to the digital divide, but its "technical configuration positions Facebook almost like an Internet service provider, allowing them to collect unique streams of user metadata" (Nothias, 2020, p. 338). The service channelled

users exclusively through Facebook's own ecosystem, treating "disadvantaged communities and less regulated territories as testing grounds for data extraction" (Nothias, 2020, p. 331). In Bangladesh, meanwhile, biometric registration of Rohingya refugees proceeded under the humanitarian rationale of identification for protection. These systems enrolled displaced populations into digital identity databases without meaningful consent. The communities subjected to these practices are "disproportionately affected by the convergence of digital developments, capitalism, and colonial legacies" (Madianou, 2019, p. 11), yet the infrastructure through which that convergence operates is rendered invisible by the vocabulary of humanitarian necessity.

The structural conditions under which corporate data extraction proceeds in humanitarian and development contexts have been theorised with increasing precision across several bodies of scholarship. The data colonialism thesis, as Couldry and Mejias (2019) formulate it, identifies a new stage of capitalism in which human life is appropriated as data and annexed to capital. The technocolonialism framework, as Madianou (2019) develops it, specifies how this logic operates within humanitarian settings through the convergence of digital innovation, market forces, and colonial dependency. Recent reviews of the field confirm that these frameworks have reshaped critical data studies (Valente and Grohmann, 2024), and the intellectual history of digital colonialism has mapped the longer trajectory of these ideas across disciplines (Nothias, 2025). What remains unspecified across this scholarship are the discursive mechanisms through which extraction is legitimated as care. Equally absent is an account of how different data practices form an interconnected dependency chain rather than operating as discrete interventions.

The political economy of datafication has been theorised as a contemporary extension of colonial resource seizure, in which the appropriation of "human life itself... so that it can be annexed directly to capital" (Couldry and Mejias, 2019, p. 2) recasts data extraction as a

political act rather than a technical process. Data functions as a form of capital that drives organisational behaviour (Sadowski, 2019, p. 2), and the humanitarian technology literature has established that digital tools reconfigure relationships of power between intervening institutions and affected communities (Sandvik et al., 2014; Jacobsen, 2015). The intellectual history of digital colonialism reinforces this account, establishing that "these technologies are new components of an old system of racialized surveillance" (Nothias, 2025, p. 385). Across these contributions, the structural architecture of data colonial extraction has been mapped, but the communicative mechanism through which that architecture is legitimated remains unexamined.

The gap is consistent across otherwise diverse bodies of scholarship. The reframing of "political problems in line with their own business objectives" by corporations that position themselves as agencies for development (Madianou, 2019, p. 5) has been identified at the level of institutional strategy, but the linguistic construction through which this reframing is accomplished remains unspecified. The discourse of "benevolence," as Nothias (2025, p. 386) establishes, is a defining feature of technological expansion into the Global South. Yet the specific vocabularies through which that benevolence is performed have not been subjected to systematic analysis. The rhetoric of decolonisation itself risks co-optation as an institutional posture that leaves power structures untouched. Dominant AI ethics frameworks have been shown to "obscure the unequal impacts of AI" on the majority world (Ricaurte, 2022, p. 729), yet the focus in that literature falls on systemic violence rather than on the legitimating vocabulary that renders violence as service. Even publicly funded North-South research collaborations struggle with subtle mechanisms of paternalism that reproduce extractive dynamics (Helm et al., 2023, p. 2). A second absence concerns the relationship between data practices. Connectivity provision, cloud hosting, data analytics, and digital

identification are studied separately across the literature, but the infrastructure dependency that connects them into a single chain of extraction has not been theorised.

This article addresses both gaps. It argues that corporate data infrastructures in humanitarian and development contexts produce a distinctive form of data colonialism in which extraction is communicatively constituted as care through specific legitimating vocabularies. The argument draws on a corpus of organisational communications, media coverage, civil society reports, and policy documents across eight cases spanning four data practices. The article makes two contributions. It provides the first systematic discourse-analytic account of how data colonial extraction is legitimated across multiple data practices. It also demonstrates that these practices form a dependency chain, with connectivity enabling cloud hosting, cloud hosting enabling analytics, and analytics enabling identification systems that close the loop of extraction.

The article proceeds as follows. Section 2 develops the theoretical framework, situating the analysis within data colonialism and critical data studies. Section 3 presents the methodology. Section 4 presents the findings integrated with discussion, organised by data practice. Section 5 concludes with implications for scholarship and governance.

2. Theoretical Framework

The argument developed in this article requires a framework that can hold together two analytical tasks. It must explain why corporate data extraction in humanitarian settings follows a colonial structural logic. It must also explain how that extraction is communicatively produced as care. Existing scholarship provides robust accounts of the first task but leaves the second largely unspecified. The framework therefore builds in sequence, moving from structural logic to discursive mechanism, from discursive mechanism to infrastructural dependency, and from dependency to the foreclosure of political contestation.

Data colonialism as structural logic

The starting claim is that contemporary data extraction operates as a form of colonial appropriation. The data colonialism thesis, as Couldry and Mejias (2019, p. 2) formulate it, identifies a new stage of capitalism in which "human life itself" is "appropriated so that it can be annexed directly to capital." This formulation recasts datafication not as a technical process of quantification but as a political act of seizure. The analogy with historical colonialism is structural rather than metaphorical. Just as colonial extraction treated land, labour, and natural resources as inputs for metropolitan accumulation, data colonialism treats the recorded traces of human experience as raw material for platform capital (Couldry and Mejias, 2019). The claim that "data is the new oil," as Couldry and Mejias (2019, p. 3) observe, naturalises this outcome by presenting extraction as the inevitable exploitation of a pre-existing resource rather than as the construction of new social relations designed to make extraction possible. The framework is powerful at the level of political economy, but its analytical register remains oriented toward capital accumulation. It does not address the question of what happens to the populations whose life chances depend on the data being extracted.

The necropolitical tradition addresses precisely this question, pushing beyond political economy into a theory of sovereignty over life and death. Where Couldry and Mejias locate the colonial logic of data extraction in the accumulation of capital, the framework that Mbembe (2019, p. 66) proposes argues that sovereign power resides in "the power and capacity to dictate who is able to live and who must die." This reframes data infrastructure as a site where that capacity is exercised. The power to determine who receives aid, and therefore who survives, through algorithmic targeting and biometric registration is a form of necropower operating through data rather than through direct violence. A related formulation in the same work identifies "domination without responsibility" as a condition in which

"capital confiscated for itself the right of life and death over those it subjugated" (Mbembe, 2019, p. 33). Corporate data actors in humanitarian settings exercise authority over the data that governs survival without accountability to the populations whose records they hold. The dependency chain analysed in this article produces precisely this condition, in which authority and responsibility are structurally separated by the architecture of the data relationship. The two frameworks are complementary but operate at different analytical registers. The question that neither answers is how this separation of authority and responsibility is communicatively sustained.

This structural logic operates at a high level of abstraction. It identifies the global architecture of data-driven capitalism but does not specify how that architecture manifests in particular institutional settings. The datafication-as-accumulation literature sharpens the political economy by demonstrating that data functions as a form of capital that drives organisational behaviour and technological development (Sadowski, 2019, p. 2). Data, in this account, is a product of social relations that have been deliberately constructed to enable continuous extraction, rather than a naturally occurring substance awaiting collection (Thatcher et al., 2016, p. 991). Accumulation by dispossession proceeds through the conversion of everyday human activity into computable, tradeable, and governable units of information.

The technocolonialism framework, as Madianou (2019) develops it, specifies this general logic within humanitarian settings. Where Couldry and Mejias theorise at the level of global capitalism and Mbembe at the level of sovereign power, Madianou identifies the specific institutional site where both logics converge. The convergence of "digital developments with humanitarian structures and market forces reinvigorates and reshapes colonial relationships of dependency" (Madianou, 2019, p. 1), operating through overlapping logics including accountability to donors, results-based management, solutionism, and the

securitisation of migration (Madianou, 2019, pp. 4–5). A subsequent formulation recasts technocolonialism as "infrastructural violence" (Madianou, 2025, pp. 179-198), shifting the analytical focus from co-occurrence to structural dependency by showing that convergence operates across layers of technology that constitute a system of control. Yet even this updated formulation identifies the fact of convergence without specifying the communicative work through which convergence is rendered acceptable.

The gap is consistent across these otherwise distinct positions. The data colonialism thesis identifies the fact of appropriation but not the communicative work through which appropriation becomes acceptable. The technocolonialism framework identifies the convergence of multiple logics but not the discursive mechanisms through which that convergence is produced and sustained. Neither framework accounts for how different data practices relate to each other as an interconnected system rather than as discrete interventions that happen to share a colonial structure.

The discursive constitution of extraction as care

The structural frameworks leave a specific gap. They explain that extraction occurs and that it reproduces colonial power relations, but they do not explain how extraction comes to be understood, accepted, and actively promoted as beneficial activity. This gap matters because legitimation is not epiphenomenal. Rather than accompanying extraction as a post hoc justification, it produces the conditions under which extraction becomes possible.

The critical humanitarianism literature provides the conceptual bridge between structural logic and discursive mechanism. The proposition that humanitarian governance "may have its heart in the right place, but it is still a form of governance, and governance always includes power," as Barnett (2011, p. 12) argues, establishes that care and power are not opposing forces but co-constitutive dimensions of humanitarian action. The legitimating

vocabularies analysed in this article are the digital form of that co-constitution. A broader historical reading reinforces the point. The emergence of an international ethics of care, as Barnett (2011, p. 8) shows, has been driven "not only by rounds of extraordinary violence but also by attempts by the international community to rescue its own self-image as civilized, humane, and good." The vocabulary of care serves the institutional self-image of the intervening actor as much as it serves the needs of the affected population. Corporate data actors inherit this dynamic when they enter humanitarian settings under the banner of digital transformation and evidence-based protection.

Where Barnett identifies the co-constitution of care and power as a structural feature of humanitarianism, Kapoor (2008, p. 1) pushes further into the postcolonial tradition, arguing that "noble gestures such as giving foreign aid or promoting participation and democracy frequently mask institutional biases and economic and geopolitical interests, while silencing the subaltern." The distinction matters. For Barnett, care and power coexist within humanitarian governance. For Kapoor, the gesture of care actively produces silencing as its structural consequence. The legitimating vocabularies documented in this article are the digital form of these noble gestures. They mask the extractive logic of the data relationship while foreclosing the discursive space within which affected populations might articulate an alternative account of what is happening to their data.

The digital colonialism literature locates these dynamics within a longer intellectual history. A discourse of "benevolence," as Nothias (2025, p. 386) establishes, is a defining feature of technological expansion into the Global South. Earlier work in the same tradition traces how this benevolence operates through a persistent pattern in which "disadvantaged communities and less regulated territories" are positioned as "testing grounds for data extraction and technological experiments" (Nothias, 2020, p. 331). The framing is constitutive of the power relation, because it determines the terms on which the population

enters the data relationship. When a population is addressed as a beneficiary of connectivity or a recipient of digital inclusion, the terms of the encounter have already been set. The population enters the data relationship as a grateful recipient of a service whose extractive dimensions are rendered invisible by the vocabulary of care, rather than as a rights-bearing political community that might negotiate or refuse extraction.

The humanitarian technology literature reinforces this account. Digital tools reconfigure relationships of power between intervening institutions and affected communities (Sandvik et al., 2014, p. 221). The reconfiguration proceeds through the centralisation of authority over data and the reduction of complex human situations to computable variables, while political questions are displaced by technical solutions (Jacobsen, 2015, pp. 2–3). The critical data studies literature demonstrates that dominant ethical frameworks developed in Western institutional contexts "obscure the unequal impacts of AI" on the majority world (Ricaurte, 2022, p. 729). These frameworks substitute procedural compliance for substantive justice, treating informed consent as a sufficient ethical safeguard even when the conditions for meaningful consent do not exist.

The analytical move this article proposes is that legitimating vocabularies are constitutive of the structural logic of data colonialism, not secondary to it. Rather than describing a relationship of care that exists independently of the data practice, they produce the subject position from which the data practice appears as care. This move departs from existing accounts of data colonialism, which tend to treat discourse as ideological overlay on a material base. The data justice literature has called for attention to the epistemic dimensions of datafication (Taylor, 2017, p. 6) and to the ways in which the "majority world" is rendered invisible by classificatory systems designed elsewhere (Ricaurte, 2022, p. 730). The argument developed here takes that call seriously by treating legitimating vocabularies as a primary site of analysis rather than as a secondary effect of structural power.

The dependency chain of data practices

Existing scholarship tends to analyse data practices in isolation, studying biometric registration as a problem of identification and surveillance, cloud infrastructure as a problem of digital sovereignty, connectivity provision as a problem of access or platform enclosure, and predictive analytics as a problem of algorithmic governance. Each of these literatures produces important findings, but the separation obscures a structural relationship. These practices form a dependency chain in which each enables and requires the others.

The dependency chain operates as follows. Connectivity infrastructures constitute populations as digital users and generate behavioural metadata. The zero-rated services studied in the digital colonialism literature channel users through proprietary ecosystems, creating what amounts to a "walled-garden" experience that produces unique streams of data unavailable through open internet access (Nothias, 2020, p. 338). Cloud infrastructure houses and processes that data under Northern corporate jurisdiction. The legal architecture governing this jurisdiction replicates a colonial pattern. The necropolitical analysis of colonial law identifies a condition in which law was "unconditionally subject to political imperatives," a conception that "worked to free power holders of any meaningful constraint" (Mbembe, 2019, p. 25). The US CLOUD Act functions analogously, subjecting humanitarian data held by American companies to US political imperatives regardless of where the data was generated or where the servers are physically located. The CLOUD Act compels disclosure of data held by American companies regardless of the physical location of the servers, meaning that humanitarian data hosted on Microsoft Azure or Amazon Web Services is legally tethered to US law enforcement authority even when it concerns populations in Bangladesh or Kenya. Analytics platforms transform the stored data into governance decisions. The partnership between WFP and Palantir shows how humanitarian data, once collected and hosted, becomes the input for analytical systems whose operational logic is

determined by the corporate provider rather than by the affected population (Madianou, 2019, p. 1). Identification systems make populations legible to all preceding layers by encoding individual identities into machine-readable formats. Biometric data, once collected, cannot be withdrawn. It locks the individual into the data relationship permanently.

The infrastructure studies literature provides theoretical grounding for this claim. The platformisation of infrastructure means that formerly public utilities are increasingly provided through proprietary corporate systems that extract value from the data flows they enable (Plantin et al., 2018, p. 298). The "smartness mandate" normalises data-intensive governance by framing the collection and analysis of population-level data as an obvious feature of modern institutional competence (Halpern and Mitchell, 2023, p. 4). Platform capitalism, understood as an economic model organised around the extraction of data from intermediary positions in social life, provides the business logic that drives corporate engagement with humanitarian settings (Srnicek, 2017, p. 43). These theoretical traditions converge on a single point. Data practices in humanitarian contexts are structurally interdependent components of a single extractive apparatus rather than discrete interventions with independent logics.

Subject positions and the foreclosure of contestation

The dependency chain produces a political consequence that extends beyond extraction itself. Each data practice constitutes affected populations in a specific subject position that determines their capacity to contest the terms of the data relationship. A connectivity user retains some navigational agency within the platform ecosystem, although that agency is structured by the proprietary architecture of the service. A person whose data is cloud-hosted retains almost no practical capacity to access, audit, or challenge the terms under which their information is stored and processed. A biometrically registered person

retains no capacity at all, because biometric data is irreversible. Once an iris scan or fingerprint is enrolled, it cannot be unenrolled.

The consent dilemma is central to this analysis, and the critical humanitarianism literature confirms its structural nature. Humanitarian actors "arrive in highly deprived environments with various privileges and resources that make any notion of consent inherently problematic" (Barnett, 2011 [p. TBC]). This problem is recognised within the humanitarian ethics tradition itself, not merely asserted by external critics. The data practices documented in this article compound the problem because digital infrastructure adds technical irreversibility to the pre-existing structural asymmetry. In every case in the corpus, data collection is framed as a condition for receiving care, aid, identity, or connectivity. Refusal means exclusion from life-sustaining services. The choice facing refugees and crisis-affected populations is structurally equivalent to what has been described as a "Hobson's choice" in which the only alternative to data extraction is the forfeiture of protection (Madianou, 2019, p. 8). "Informed consent" is structurally impossible when the alternative to consent is hunger, statelessness, or invisibility. This dynamic is particularly evident in the case of Rohingya biometric registration. Refugees were enrolled into identity databases as a condition of receiving food assistance, without meaningful opportunity to refuse or to negotiate the scope of data collection (Human Rights Watch, 2021, p. 3). The data collected was subsequently shared with actors whose interests diverged from those of the registered population.

The consent dilemma connects back to the legitimating vocabularies identified in the second layer of this framework. Rather than describing the data relationship, the vocabulary of care produces the subject position from which affected populations cannot contest its terms. When extraction is framed as "protection" or "inclusion," refusal becomes legible only as self-harm. A person who refuses biometric registration is understood as failing to access a

service designed for their benefit, rather than as exercising a political right to data sovereignty. The discursive construction of extraction as care thus forecloses the political space within which contestation might occur.

The epistemic dimension of this foreclosure has deeper roots in the history of colonial knowledge production. The argument that Western civilisation "managed to have the epistemic privilege of narrating its own local history and projecting it onto universal history" (Mignolo, 2000/2012, Preface) identifies the mechanism through which particular knowledge systems are naturalised as universal. A related argument establishes how modernity itself is constituted through the suppression of alternative epistemologies, producing a global order in which Western categories appear as the only available framework for understanding (Mignolo, 2011). Northern data standards, biometric categories, and algorithmic classifications replicate this projection. They encode particular assumptions about identity, risk, and need into technical systems that are then deployed as if they were neutral instruments of measurement. The subject positions produced by the dependency chain are therefore not only political but epistemic. Affected populations are foreclosed not only from contesting the terms of the data relationship but from articulating alternative frameworks within which data might be understood.

This is the mechanism through which data colonialism is communicatively sustained. The structural logic identified by the data colonialism thesis produces the conditions for extraction. The legitimating vocabularies produce the conditions under which extraction appears as care. The dependency chain connects discrete data practices into a single system. The subject positions produced by that system foreclose the possibility of refusal. The framework developed here holds these four dimensions together as a single analytical architecture, which the empirical analysis applies to eight cases spanning four data practices.

[Figure 1 about here]

3. Methodology

Research design

I adopted a qualitative multiple-case study design (Stake, 1995; Yin, 2018) with discourse-analytic coding categories applied through a framework analysis procedure (Ritchie and Spencer, 1994). The case study design provides the sampling and comparison structure. The discourse-analytic categories, drawing on the interpretivist tradition of analysing how language constitutes social relations and subject positions, specify what is analysed within that structure. Framework analysis provides the organising method, moving from theoretically prior categories through systematic coding to cross-case interpretation while accommodating categories that emerge during analysis.

The unit of analysis is the data practice rather than the individual case. I organised seven cases across four data practices (analytics, connectivity, cloud infrastructure, and identification), with one contrast case that falls outside the dependency chain. The multiple-case design enables comparison at two levels. Within each data practice, I compared cases to identify the shared legitimating vocabulary and its variations. Across data practices, the comparison tests whether the dependency chain theorised in the framework operates empirically. The collective case study approach treats each case as instrumental to understanding the broader phenomenon rather than as intrinsically bounded (Stake, 1995).

Data collection

I selected cases through purposive sampling to achieve maximum variation across four dimensions. These were data practice (analytics, connectivity, cloud, and identification), corporate actor type (technology companies and multilateral organisations), geographic scope (cases spanning multiple Global South contexts rather than single-country studies), and the

presence of documented contestation. The requirement for documented contestation is analytically motivated. The article's argument concerns how extraction is legitimated as care. That argument can only be demonstrated where the legitimating vocabulary is visible in organisational communications and where the counter-discourse is visible in contestation documents.

The four data practice categories are derived from the theoretical framework's dependency chain. The case selection therefore operationalises the framework's central claim that these practices form an interconnected system of extraction rather than discrete interventions. I included an eighth case, Sama/Meta, as a contrast case. Content moderation labour in Kenya does not belong to any of the four data practices in the dependency chain. It operates through direct labour exploitation rather than infrastructural dependency. The contrast case tests the framework's boundaries by examining whether the legitimating vocabulary identified across the dependency chain also operates in a data practice whose extractive logic is organised differently. Table 1 presents the eight cases and their assignment to data practices.

The empirical material comprises approximately 170 publicly available documents collected across the eight cases. For each case, I collected four document types, each serving a distinct analytical function. Organisational communications (press releases, programme descriptions, partnership announcements, and executive statements) are the primary texts through which the legitimating vocabulary is produced. Media coverage (news reporting, investigations, and editorials) records public reception, ratification, and contestation. Civil society reports (advocacy reports, open letters, and legal filings) provide the counter-discourse against which the legitimating vocabulary can be tested. Policy and governance documents (internal audits, regulatory decisions, and parliamentary testimony) provide the institutional and regulatory context within which the data practice operates.

I collected documents through systematic web search, institutional press room review, and civil society organisation archives. I continued collection until theoretical saturation was reached for each data practice (Glaser and Strauss, 2017). The restriction to English-language sources privileges globally visible actors and may exclude contestation documented in other languages.

Data analysis

The coding categories draw on the principle, established in critical discourse analysis, that discourse is constitutive of social identities, social relations, and systems of knowledge rather than merely reflective of them (Fairclough, 1992). The organizational discourse analysis tradition specifies how this constitutive function can be examined empirically through systematic comparison of texts produced by different institutional actors about the same practices (Phillips and Hardy, 2002). The concept of subject position, as developed in the discursive positioning literature (Davies and Harré, 1990), holds that discourse produces the positions from which individuals can speak and act, determining what forms of agency are available within a given discursive frame. I coded each document along four dimensions, each operationalising a concept from the theoretical framework. Table 2 summarises the dimensions and the analytical operation applied to each.

I followed four analytical stages. The first was familiarisation with the corpus and indexing of extracts against the coding dimensions. The second was charting of indexed material into case summary templates. The third was cross-practice mapping, comparing cases within each data practice and then across data practices to test the dependency chain. It was during this third stage that the structural persistence finding emerged. The lock-in mechanisms coded under the dependency chain dimension revealed a temporal pattern that the theoretical framework had not anticipated, with no data practice in the corpus having been

dismantled after its activating rationale expired. Framework analysis accommodates such inductive emergence because it is designed to work with both prior and emergent categories (Ritchie and Spencer, 1994). The fourth stage was interpretive synthesis. I developed and tested the analytical procedure in a companion study (Anonymised, under review), but the description provided here is sufficient for the article to stand alone.

[Figure 2 about here]

To test my interpretations, I read each case's organisational communications against its media coverage and civil society contestation, checking whether the legitimating vocabulary I identified in corporate texts was recognised, ratified, or challenged in the other document types. Convergence across document types strengthens interpretive confidence. Divergence is treated as analytically informative evidence of contested legitimation. I maintained a structured analytical memo record throughout coding to track evolving interpretations and to make my reasoning auditable. The full empirical corpus is deposited at Harvard Dataverse so that other researchers can inspect the documents on which the analysis rests (Lincoln and Guba, 1985).

The corpus is publicly available at [DOI]. Full-text documents are provided where copyright permits. Metadata records with original URLs are provided for copyrighted sources.

4. Discussion

This article asked how corporate data infrastructures in humanitarian and development contexts legitimate extraction as care. The analysis of approximately 170 documents across eight cases produces a clear answer. The legitimation operates through specific vocabularies that vary in register across data practices but perform the same function

in every case. Whether the register is "evidence-based protection" in predictive analytics, "connecting the next billion" in connectivity, "sovereign cloud" in cloud infrastructure, or "legal identity for all" in digital identification, the vocabulary constitutes a data relationship as morally necessary. It defines communities as recipients of care rather than as agents who might contest its terms. Four propositions develop this argument. The data practices are not discrete interventions. They form a dependency chain that routes data from Global South populations to Northern corporate infrastructure while the vocabulary of care travels in the opposite direction. Once activated under crisis or development rationales, the infrastructure persists through lock-in mechanisms that outlast any continuing justification. The communities whose data is extracted are absent from the organisations that contest the extraction.

The section develops this argument through four propositions. The first (4.1) establishes that the legitimating vocabulary is constitutive rather than descriptive. The second (4.2) traces the dependency chain as geographical infrastructure. The third (4.3) addresses structural persistence, demonstrating that crisis functions as an activation mechanism for permanent infrastructure. The fourth (4.4) examines the consent dilemma and the geography of accountability. The contrast case (Case 08, Sama-Meta) is addressed within sections 4.1 and 4.3, where it confirms the discursive finding while limiting the structural one. Table 1 maps how affected communities appear in the contestation record. Table 2 summarises the analytical findings by data practice.

4.1 The legitimating vocabulary is constitutive, not descriptive

The data colonialism thesis, as Couldry and Mejias (2019, p. 14) develop it, identifies a structural logic in which datafication "degrades" the social experience it claims to optimise. That framework operates at a high level of abstraction. It names the fact of extraction without

specifying the communicative mechanism that makes extraction legible as care. The evidence presented here suggests that legitimating vocabularies are not secondary to the structural logic of data colonialism. Because the same foreclosure of contestation was observed across all four registers, the findings indicate that the vocabulary is the mechanism through which that logic becomes operational.

Table 2 documents variation in legitimating register across the four data practices, with analytics deploying epistemic authority, connectivity deploying inclusion, cloud infrastructure deploying innovation, and identification deploying protection. The analytical question is whether this variation matters. The evidence indicates that it does not, because different registers produce the same outcome. As Table 2 shows, each register assigns affected communities a subject position from which the terms of the data relationship cannot be negotiated. The critical humanitarianism literature establishes why. As Barnett (2011, p. 12) maintains, humanitarian governance "may have its heart in the right place, but it is still a form of governance, and governance always includes power." The data practices documented here specify the mechanism through which that power is reproduced in digital form.

The gap between public and internal documents in the Project Nimbus case (Case 02) reveals the constitutive function with particular clarity. The Google Cloud Blog announcement framed the contract through public-service modernisation, referencing healthcare and education, with no mention of defence applications. A leaked internal risk assessment, however, showed analysts acknowledging that military use could not be controlled. The vocabulary was not an afterthought applied to a contract already determined by commercial logic. It was the precondition that made the contract publicly presentable.

This constitutive function extends the technocolonialism framework. The concept of "imperial formations," as Madianou (2019, p. 7) develops it, identifies how digital practices in the Global South materialise colonial logics through humanitarian structures. The

convergence thesis that Madianou advances establishes that these logics co-occur. The evidence presented here specifies the communicative mechanism through which convergence is produced and sustained. The Palantir-WFP partnership (Case 01) is a clear instance of this. The vocabulary of "data-driven efficiency" did not describe a pre-existing relationship of care. Rather than accompanying convergence, it produced the conditions under which the transfer of 90 million aid recipients' data to a company simultaneously serving US immigration enforcement became legible as humanitarian innovation.

The Free Basics case (Case 03) reveals that the vocabulary's effectiveness depends on conditions external to its content. Meta used the same language in India and across 32 African countries. In India, an organised civil society and an assertive telecommunications regulator blocked the service. What Nothias (2020, p. 332) terms "benevolent capitalism" failed where contestation infrastructure existed and succeeded where it did not. Contestation capacity is therefore shaped by the same structural inequalities that the data practice exploits.

The intellectual history of digital colonialism, as Nothias (2025, p. 388) traces it, identifies a persistent pattern of "malevolent paternalism" in which technological solutions are offered for the perceived good of colonised populations. The evidence developed here goes further. The legitimating vocabularies actively produce the conditions under which what Ricaurte (2022, p. 727) calls "North-South systemic asymmetries" are rendered invisible, rather than merely deepening them. The vocabulary of "evidence-based protection" constitutes the data relationship as inherently protective, such that harm becomes conceptually unavailable within the terms of the discourse. The contrast case (Case 08, Sama-Meta content moderation) confirms the pattern. Kenyan moderators reviewed graphic material for approximately $2.20 per hour under the banner of "ethical AI" and "dignified digital work." When workers organised, they were dismissed. The discursive mechanism

operates even where the extractive relationship runs through employment contracts rather than through data infrastructure.

4.2 The dependency chain as geographical infrastructure

Existing scholarship tends to analyse data practices in isolation. Biometric registration, cloud sovereignty, platform connectivity, and predictive analytics occupy separate literatures with limited cross-citation (Jacobsen, 2015; Sandvik et al., 2014). The zero-rated services that function as "walled gardens" (Nothias, 2020, p. 330) are studied apart from the cloud infrastructures that host the data those gardens generate. The dependency chain concept demonstrates that these practices are infrastructurally connected, and that the connection is geographical.

The Palantir-WFP partnership (Case 01) demonstrates the pipeline operating end to end. The WFP SCOPE biometric database, containing records on over 90 million aid recipients, was created through identification (Cases 06, 07). That data is hosted on corporate cloud infrastructure under US jurisdiction (Case 05). Palantir's analytics platform then processes it into targeting decisions, while the same company simultaneously held enforcement contracts with US Immigration and Customs Enforcement. A study of any single data practice would see only one segment. The cross-practice design makes the full routing visible, from biometric enrolment in the Global South to analytical processing under US jurisdiction. The Project Nimbus case (Case 02) adds a further dimension. The analysis of necropolitics identifies occupied Palestine as "a laboratory" for techniques of "control, surveillance, and separation" that subsequently proliferate elsewhere (Mbembe, 2019, pp. 42–43). The dependency chain documented here traces the infrastructural pathway through which that proliferation operates.

Across all eight cases, the analysis revealed a single directional structure. This is perhaps the most striking finding to emerge from the cross-practice comparison. It is particularly clear in the case of the CLOUD Act (Case 05), where data generated by populations in Bangladesh or Kenya is legally tethered to US jurisdiction, and in the Palantir partnership (Case 01), where analytical authority over 90 million aid recipients' records is exercised from a US corporate headquarters. Data flows from Global South populations to Northern corporate infrastructure, where authority over that data is exercised and where the capital it generates accumulates. The legitimating vocabulary flows in the opposite direction, from North to South, constituting the extraction as care. This finding is broadly consistent with the platformisation thesis advanced by Plantin et al. (2018, p. 298), who argue that formerly public utilities are increasingly provided through proprietary corporate systems that extract value from the data flows they enable. It suggests, however, that data colonialism operates as a geographical hierarchy rather than as convergence in the abstract sense, with the discursive and material dimensions moving in opposite directions.

A possible explanation for this geographical consistency across otherwise diverse cases is that the normalisation of data-intensive governance as an obvious feature of modern institutional competence (Halpern and Mitchell, 2023, p. 4) creates a structural demand for corporate data infrastructure that only Northern providers can supply at scale. The dependency is infrastructural before it is political. By the time the political consequences become visible, vendor lock-in and jurisdictional tethering are already in place.

4.3 Structural persistence and the function of crisis

An unexpected finding to emerge from the analysis is its temporal dimension, which the theoretical framework did not anticipate. None of the eight data practices has been dismantled after its original justification expired. The Rohingya biometric registration (Case

06), initiated during the 2017 displacement, entered its ninth year in 2026. In 2025, families who refused renewed biometric enrolment were cut from food assistance. The crisis rationale is nine years old, yet the infrastructure continues to demand compliance as a condition of survival. The same persistence is visible across the other cases. The Free Basics service (Case 03) and the ID4D programme (Case 07) have operated for thirteen and ten years respectively, while the CLOUD Act (Case 05) contains no sunset clause and the Project Nimbus contract (Case 02) includes a 23-year extension option with an anti-boycott provision.

This pattern suggests that crisis is not merely the context in which these data practices are deployed. It functions as the initial justification that enables permanent infrastructure. The historical pattern is familiar. When colonial authorities absorbed relief functions, "shifting the locus of relief to the state effectively gave the British colonial authorities more power over the populations, though without any appreciable improvement in their welfare" (Barnett, 2011, p. 64). The Rohingya case (Case 06) is a notable example of this logic in digital form. The 2017 displacement activated the biometric registration, but the infrastructure persists nine years later, demanding compliance as a condition of food assistance. Because none of the eight practices has been dismantled after its rationale expired, the evidence suggests that, once operational, the infrastructure no longer requires the vocabulary that enabled it. The convergence of digital developments and market forces does not simply turn inequalities into "hard facts" that are difficult to challenge (Madianou, 2019, p. 11). It renders them irreversible through technical, legal, and contractual mechanisms that outlast any continuing justification.

The contrast case qualifies this finding. When Sama ended its Meta contract in 2023, Meta contracted other subcontractors. Lock-in in the data labour case is economic rather than technical, and a precarious worker retains some capacity to organise, as the named plaintiff Daniel Motaung demonstrated through Foxglove. A biometrically registered refugee has no

equivalent mechanism because biometric data cannot be withdrawn. The pipeline and permanence findings therefore apply to the seven infrastructural cases but not to data labour.

4.4 The consent dilemma and the geography of accountability

If the vocabulary forecloses contestation from within the data relationship, and if the infrastructure outlasts its justification, who contests the consent dilemma, and from where? Table 1 answers this question case by case. In every instance, the organisations that document and challenge the consent dilemma are located outside the communities that experience it. Affected communities appear not as agents who can alter the data relationship but as subjects of witnessing, present through the accounts that external organisations produce about them (Peters, 2001, p. 710). The postcolonial critique of development identifies a deeper problem. The intention behind intervention, whether conscious or not, may be "to control or appropriate" rather than to empower (Kapoor, 2008, p. 199). The data practices documented here operationalise that logic through technical infrastructure rather than through institutional paternalism alone.

The evidence of consent failures in the Rohingya case (Case 06) comes from a Human Rights Watch investigation conducted from New York. The evidence of ID4D exclusion (Case 07) comes from a NYU report, also produced in New York. Refugees, users, citizens, and workers enter the evidentiary record only through external mediation. The observation that beneficiary voices "have been largely absent from both scholarly and NGO discussions" (Seu and Orgad, 2017, p. 128) applies with equal force to the data governance literature.

If consent is structurally impossible when data collection is a condition of survival, then consent-based data governance frameworks are insufficient (Taylor, 2017, p. 6). The critical humanitarianism literature confirms the structural nature of this problem. Humanitarian actors "arrive in highly deprived environments with various privileges and

resources that make any notion of consent inherently problematic" (Barnett, 2011 [p. TBC]). Corporate data actors operate under the same asymmetry, compounded by technical irreversibility. The Rohingya case is a case in point. The families cut from food assistance in 2025 for refusing biometric re-enrolment had adequate information about the registration process. What they lacked was the structural capacity to refuse without forfeiting access to sustenance. This finding is consistent with the necropolitical framework, which specifies what such a condition means for the populations concerned. They are rendered into what the literature describes as people who "live at the edge of life" (Mbembe, 2019, p. 37), governed through a data relationship they cannot exit.

The geography of accountability is inverted. Accountability runs upward to Northern donors and technology partners, not downward to the Southern populations whose data is extracted. Another important finding is that in no case does contestation originate from the affected community itself. The observation that digital tools reconfigure relationships of power between intervening institutions and affected communities (Sandvik et al., 2014, p. 221) is confirmed by this corpus. The reconfiguration is geographical. It transfers authority over data from the location where it was generated to the jurisdiction where it is stored and processed. This pattern may reflect the combined effect of two conditions. The first is that the communities at the extractive end of the pipeline lack the institutional infrastructure, legal standing, and financial resources to mount formal contestation. The second is that the legitimating vocabulary itself renders contestation conceptually difficult, because refusal of "protection" or "inclusion" cannot easily be articulated as a political demand within the terms the vocabulary makes available. The Microsoft French Senate testimony of 2025 illustrated the jurisdictional dimension. The company admitted under oath that it cannot guarantee European data will remain protected from US government access. This admission suggests

that if sovereignty cannot be guaranteed for European governments, it is unlikely to be guaranteed for the far less powerful populations of the Global South.

Limitations and future research

The present analysis is not exempt from the asymmetry it documents. The documentary corpus consists of Northern investigative reports, Northern legal advocacy, and Northern-affiliated academic fieldwork. This asymmetry is a structural condition of the evidentiary base available for this kind of analysis, and it specifies the epistemic boundaries within which the findings operate. The corpus is restricted to English-language sources. This restriction privileges Northern media and Northern civil society perspectives. The voices of affected populations are present only as mediated through Northern documentary practices. Future research must address this asymmetry directly through participatory methods and Southern-led research partnerships. The set of political practices that the literature terms "algoritharism" (Ricaurte, 2022, p. 732), through which technological arrangements shape futures under conditions of domination, can only be contested from within the communities it governs.

The dependency chain concept requires further empirical testing across cases and regions not included in this corpus, including Latin America and East Asia. The structural persistence finding requires longitudinal investigation to determine whether and how the data practices documented here evolve as regulatory environments shift. The contrast case suggests that data labour is a geographical practice in the chain that this article could only begin to examine.

5. Conclusion

This article set out to identify the communicative mechanism through which corporate data extraction in humanitarian and development contexts is legitimated as care, and to determine whether the data practices through which this extraction operates form an interconnected system. The analysis indicates that data extraction in these contexts is communicatively constituted as care. The legitimating vocabulary varies by data practice, ranging from "evidence-based protection" in analytics to "connecting the next billion" in connectivity, from "sovereign cloud" in cloud infrastructure to "legal identity for all" in identification. The register changes, but the constitutive operation is constant. In every case, the vocabulary constitutes a data relationship as morally necessary and forecloses the subject position from which affected communities could contest its terms. The four data practices form a dependency chain that operates as a geographical pipeline, routing the vocabulary of care from North to South while extracting value in the opposite direction.

If the argument developed here is accepted, two consequences follow. Because the analysis demonstrated that different legitimating registers produce the same foreclosure of contestation across all four data practices, it follows that structural accounts of data colonialism are insufficient on their own. The discursive dimension must be addressed alongside the material and jurisdictional dimensions, since it is the mechanism through which the structural logic becomes operational. Because the analysis also demonstrated that the four data practices are infrastructurally connected along a geographical axis, it follows that regulation operating within a single jurisdiction or targeting a single data practice will not dismantle the system. The regulatory record confirms this. The Indian TRAI ruling banned Free Basics inside India but left 32 African countries unaffected, just as European GDPR fines penalised facial recognition companies within Europe without preventing US government adoption of the same technologies. The chain persists because regulatory capacity is itself geographically distributed.

The structural persistence of every data practice in the corpus reinforces this point. Emergency framing activates the infrastructure, and contractual and jurisdictional lock-in mechanisms ensure it persists long after the activating rationale has expired. Across the corpus, data practices have operated for between nine and thirteen years, with some carrying no sunset clause and others including multi-decade extension options. None has been dismantled. The infrastructure that crisis justified has become permanent, built under the vocabulary of care and sustained under the logic of extraction.

A data justice framework adequate to these findings would need to meet two conditions. It would need to address the dependency chain as a geographically distributed system rather than as a set of local problems amenable to local regulation. It would also need to create governance structures in which communities in the Global South have standing to contest the terms of data relationships, rather than relying on Northern civil society to contest those relationships on their behalf. Such a framework would require moving beyond consent-based governance, which assumes a subject who can refuse, toward accountability structures anchored at the site of extraction rather than at the site of corporate headquarters.

The findings carry additional implications for AI ethics governance. Dominant AI ethics frameworks rest on procedural safeguards, with informed consent, transparency, and accountability operating as the primary mechanisms for ensuring that data-driven systems respect the rights of those they affect. The analysis presented here identifies a structural limitation in this approach. Consent-based governance presupposes a subject who can refuse. The vocabularies documented across the corpus produce precisely the opposite condition, in which refusal of "protection" or "inclusion" is legible only as self-harm. When the alternative to data collection is exclusion from food assistance, medical care, or legal identity, consent is not a meaningful ethical safeguard. It is a procedural formality that legitimates extraction. A second limitation concerns the unit of ethical analysis. AI ethics review typically evaluates

individual applications, assessing whether a particular algorithm is fair, transparent, or accountable. The dependency chain documented in this article demonstrates that ethical evaluation at the application level misses the systemic character of extraction. Each component in the chain may satisfy procedural ethical requirements while the system as a whole operates as an extractive apparatus. An ethics adequate to this finding would need to move from application-level review to infrastructure-level accountability, assessing not only whether individual tools meet ethical standards but whether the connected system of which they form part reproduces colonial patterns of extraction.

The findings also carry implications for how data colonialism and technocolonialism are theorised. Notwithstanding the limitations of a documentary corpus restricted to English-language sources and mediated through Northern institutional perspectives, the cross-practice comparison revealed that the same constitutive function operates through different registers. This finding suggests that the data colonialism thesis requires geographical specification if it is to move from structural diagnosis toward explanatory analysis of how extraction operates in particular settings. Because the convergence identified in the technocolonialism literature was shown to depend on specific legitimating vocabularies, that framework requires communicative specification if it is to explain how convergence is produced rather than merely identifying that it occurs. Taken together, these findings suggest that the legitimating vocabulary is the mechanism that existing accounts have left unspecified. This article contributes the first systematic discourse-analytic account of that mechanism across multiple data practices in humanitarian and development settings. The analysis presented here indicates that the vocabulary of care is the means through which extraction becomes operational, and that dismantling data colonial infrastructure requires confronting the discursive apparatus that sustains it. Future research should test the dependency chain

concept across regions not covered here, and should develop participatory methods capable of centring the voices of affected communities in the analysis of data governance.

References

Amnesty International (2020) *Failing to Do Right: The Urgent Need for Palantir to Respect Human Rights*. London: Amnesty International. AMR 51/3124/2020.

Barnett M (2011) *Empire of Humanity: A History of Humanitarianism*. Ithaca, NY: Cornell University Press.

Couldry N and Mejias UA (2019) *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford, CA: Stanford University Press.

Davies B and Harré, R. (1990) 'Positioning: The Discursive Production of Selves', *Journal for the Theory of Social Behaviour*, 20(1): 43–63.

Fairclough N (1992) *Discourse and Social Change*. Cambridge: Polity Press.

Glaser BG and Strauss AL (2017) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. London: Routledge. (Originally published 1967.)

Halpern O and Mitchell R (2023) *The Smartness Mandate*. Cambridge, MA: MIT Press.

Helm P, de Götzen A, Cernuzzi L, et al. (2023) Diversity and Neocolonialism in Big Data Research: Avoiding Extractivism While Struggling with Paternalism. *Big Data & Society* 10(2): 1–14.

Human Rights Watch (2021) *"An Island Jail in the Middle of the Sea": Rohingya Refugees on Bhasan Char*. New York: Human Rights Watch.

Jacobsen KL (2015) *The Politics of Humanitarian Technology: Good Intentions, Unintended Consequences and Insecurity*. London: Routledge.

Kapoor I (2008) *The Postcolonial Politics of Development*. London: Routledge.

Lincoln YS and Guba EG (1985) *Naturalistic Inquiry*. Newbury Park, CA: Sage.

Madianou M (2019) 'Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises', *Social Media + Society*, 5(3): 1–13.

Madianou M (2025) *Technocolonialism: When Technology for Good Is Harmful*. Cambridge: Polity Press.

Mbembe A (2019) *Necropolitics*. Durham, NC: Duke University Press.

Mignolo WD (2000/2012) *Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking*. Princeton, NJ: Princeton University Press.

Mignolo WD (2011) *The Darker Side of Western Modernity: Global Futures, Decolonial Options*. Durham, NC: Duke University Press.

Nothias T (2020) 'Access Granted: Facebook's Free Basics in Africa', *New Media & Society*, 22(2): 329–348.

Nothias T (2025) An Intellectual History of Digital Colonialism. *Journal of Communication* 75(5).

Peters JD (2001) 'Witnessing', *Media, Culture & Society*, 23(6): 707–723.

Phillips N and Hardy C (2002) *Discourse Analysis: Investigating Processes of Social Construction*. Thousand Oaks, CA: Sage.

Parker B (2019) New UN Deal with Data Mining Firm Palantir Raises Protection Concerns. *The New Humanitarian*, 5 February.

Plantin J-C, Lagoze C, Edwards PN and Sandvig C (2018) 'Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook', *New Media & Society*, 20(1): 293–310.

Ricaurte P (2022) 'Ethics for the Majority World: AI and the Question of Violence at Scale', *AI & Society*, 37: 725–735.

Ritchie J and Spencer L (1994) 'Qualitative Data Analysis for Applied Policy Research', in Bryman A and Burgess RG (eds) *Analysing Qualitative Data*. London: Routledge, pp. 173–194.

Sadowski J (2019) 'When Data Is Capital: Datafication, Accumulation, and Extraction', *Big Data & Society*, 6(1): 1–12.

Sandvik K B., Jumbert M G., Karlsrud J and Kaufmann M (2014) 'Humanitarian Technology: A Critical Research Agenda', *International Review of the Red Cross*, 96(893): 219–242.

Seu IB and Orgad S (2017) *Caring in Crisis? Humanitarianism, the Public and NGOs*. Basingstoke: Palgrave Macmillan.

Srnicek N (2017) *Platform Capitalism*. Cambridge: Polity Press.

Stake RE (1995) *The Art of Case Study Research*. Thousand Oaks, CA: Sage.

Taylor L (2017) 'What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally', *Big Data & Society*, 4(2): 1–14.

Thatcher, J., O'Sullivan D and Mahmoudi D (2016) 'Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data', *Environment and Planning D: Society and Space*, 34(6): 990–1006.

Valente JCL and Grohmann R (2024) Critical Data Studies with Latin America: Theorizing beyond Data Colonialism. *Big Data & Society* 11(1): 1–14.

Wright K (2018) *Who Is Reporting Africa Now? Non-Governmental Organizations, Journalists, and Multimedia*. Oxford: Peter Lang.

Yin RK (2018) *Case Study Research and Applications: Design and Methods*. 6th edn. Thousand Oaks, CA: Sage.

Tables

Table 1. Case selection matrix by data practice, corporate actor, and context

Data practice	Case	Corporate/institutional actor	Context
Analytics	Palantir–WFP/UNHCR	Palantir Technologies	Humanitarian data partnership
Analytics	Google/Amazon Project Nimbus	Google, Amazon	Cloud and AI for Israeli government
Connectivity	Meta Free Basics	Meta/Facebook	Internet access in India and Africa
Connectivity	SpaceX Starlink	SpaceX	Satellite internet in Ukraine
Cloud infrastructure	Microsoft Azure / CLOUD Act	Microsoft, AWS	Humanitarian data sovereignty
Identification	UNHCR Rohingya biometrics	UNHCR, IrisGuard	Refugee registration in Bangladesh
Identification	World Bank ID4D / Aadhaar	World Bank, UIDAI	Digital identity systems
Contrast (data labour)	Sama/Meta	Sama, Meta	Content moderation in Kenya

Table 2. Coding dimensions and analytical operations

Coding dimension	What was recorded	Analytical operation
1. Legitimating vocabulary	Recurrent terms and phrases framing extraction as beneficial, classified by register (care, inclusion, efficiency, protection, innovation)	Compared vocabulary in organisational communications against civil society and media accounts of the same practice, testing whether terms function descriptively or constitutively (Fairclough, 1992)
2. Subject position production	How the vocabulary addresses affected populations (as beneficiary, user, data subject, or rights-bearing agent)	Assessed what forms of agency or contestation each position makes available or forecloses (Davies and Harré, 1990)
3. Dependency chain indicators	Infrastructure reliance, data flow directionality, jurisdictional exposure, lock-in mechanisms	Traced connections between data practices across the corpus
4. Contestation dynamics	How civil society and media sources challenge the legitimating vocabulary and on what grounds	Tested whether the vocabulary identified in corporate texts was recognised, ratified, or challenged in other document types

Table 1. How affected communities appear in the contestation record

Case	Affected community	Organisation that challenged the data practice	Location of challenger	How affected community appears in the record
01 Palantir–WFP data partnership (Palantir, WFP)	Refugees across crisis zones (90M+)	Privacy International; Access Now	London; New York	Aggregate statistics; no direct testimony
02 Project Nimbus contract (Google, Amazon, Israeli govt)	Palestinian communities	No Tech for Apartheid; Abolitionist Law Center	US; US	Named as affected; represented by tech employees and legal advocates
03 Free Basics zero-rated internet (Meta)	Users in 32 African countries	Save the Internet (India only)	India	Indian users represented; African users absent from record
04 Starlink satellite internet (SpaceX)	Ukrainian citizens and military	Belfer Center, Harvard	US	Citizens appear as infrastructure dependents
05 Azure cloud hosting, CLOUD Act (Microsoft, AWS)	UN agencies; Global South data subjects	European Parliament; French Senate	Brussels; Paris	Institutional actors contest; individual data subjects absent
06 Rohingya biometric registration (UNHCR, IrisGuard)	880,000 Rohingya refugees	Human Rights Watch; Access Now	New York; New York	Interview testimony mediated through HRW
07 ID4D digital identity programme (World Bank, UIDAI)	1.2B Indian citizens; enrollees in 49 countries	NYU CHRGJ; Access Now coalition	New York; New York	Exclusion cases cited; no direct voice
08 Content moderation outsourcing (Sama, Meta)	Kenyan content moderators	Foxglove	London	Named plaintiff (Daniel Motaung); workers appear through UK legal advocacy

Table 2. Analytical summary by data practice

Data practice	Legitimating register	Subject position produced	Consent structure	Lock-in mechanism
Analytics (Cases 01, 02)	Epistemic authority ("evidence-based," "data-driven," "digital transformation")	Data subject; no governance rights over secondary uses	Consent delegated to institutional partner; individual consent absent	Multi-year partnership; anti-boycott clause; 23-year extension
Connectivity (Cases 03, 04)	Inclusion ("connecting the next billion," "humanitarian gesture")	Platform user inside walled garden	Terms-of-service acceptance; no alternative when "free" is only option	Platform dependency; sovereign dependency on single provider
Cloud infrastructure (Case 05)	Innovation ("sovereign cloud," "AI for Humanitarian Action")	Jurisdictional subject; governed by US law regardless of server location	No direct consent; institutional decision by host agency	Vendor lock-in; CLOUD Act permanent legislation
Identification (Cases 06, 07)	Protection ("documentation," "legal identity for all")	Biometric subject; irrevocable format	Registration as condition of food, shelter, legal status	Biometric data irrevocable;

				mandatory for rations, banking, tax
Data labour, contrast (Case 08)	Ethical employment ("ethical AI," "dignified digital work")	Precarious worker; economic lock-in	Accept conditions or lose employment	Labour replaceable; economic precarity

Figures

Figure 1. From extraction to foreclosure

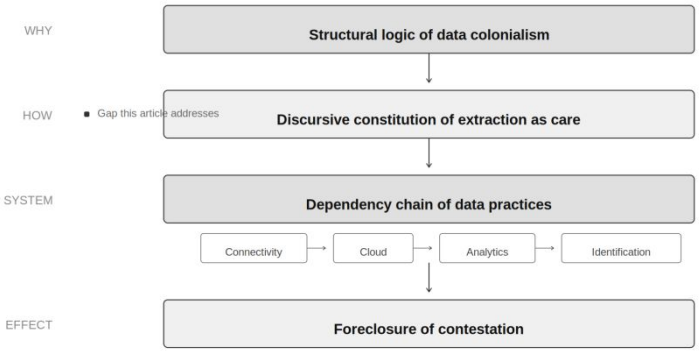

Figure 2. Analytical procedure