MAK HILL Publications International Journal of Soft Computing

Review Article
doi: 10.36478/makijsc.2024.1.8

OPEN ACCESS

**Key Words**
Artificial intelligence, benchmark datasets, data sovereignty, decolonial theory, machine learning, epistemic justice, indigenous data governance, data pluriverses

**Corresponding Author**
Erwin L. Rimban,
Department of Cagayan State University Andrews Campus Republic, Philippines

**Author Designation**
Assistant Professor

**Received:** 05th February 2024
**Accepted:** 10th March 2024
**Published:** 29th April 2024

**Citation:** Erwin L. Rimban, 2024. Data Sovereignty and the Myth of the Universal Dataset: A Critical Review of Benchmarking in Machine Learning. Int. J. Soft Comput., 19: 1-8, doi: 10.36478/makijsc.2024.1.8



# Data Sovereignty and the Myth of the Universal Dataset: A Critical Review of Benchmarking in Machine Learning

Erwin L. Rimban
*Department of Cagayan State University Andrews Campus Republic, Philippines*

**Abstract**
This paper presents a critical review of the concept of universal benchmark datasets in machine learning through the lens of data sovereignty and decolonial theory. While benchmark datasets like Image Net, COCO and GLUE have become standard tools for evaluating model performance, they often reflect Western cultural norms, linguistic biases, and geopolitical priorities. Drawing on theoretical frameworks from Walter Mignolo's epistemic disobedience, Boaventura de Sousa Santos's epistemologies of the South, Miranda Fricker's epistemic injustice and Philip Alston's digital colonialism, this paper critically examines the historical development, construction politics and universality claims of benchmark datasets. The analysis reveals how these datasets marginalize non-Western knowledge systems and perpetuate colonial power dynamics in data practices. As alternatives, this paper proposes data pluriverses, co-design frameworks for localized benchmarking, decentralized dataset stewardship and integration of Indigenous data governance principles like CARE (Collective Benefit, Authority to Control, Responsibility, Ethics). The paper concludes by emphasizing the urgent need to dismantle universalist assumptions in AI development and calls for more ethical and pluralistic data practices in machine learning research.

## INTRODUCTION

The explosive growth of machine learning (ML) across global contexts has transformed industries, governance and everyday life. This transformation has been fueled by increasingly complex models that promise to understand, predict and even generate human-like responses to visual, textual and multimodal inputs. At the center of this technological revolution lies a critical yet often overlooked foundation: the benchmark dataset. These curated collections of data have become the touchstones by which progress in the field is measured, careers are built and billions of dollars in research funding are allocated[1]. Yet beneath their seemingly neutral surface lies a complex web of cultural assumptions, epistemic frameworks and power dynamics that fundamentally shape what kinds of knowledge are valued, whose perspectives are privileged and which applications are prioritized. The term "universal dataset" refers to collections of data that purport to represent general, broadly applicable instances of a particular domain-images that represent "visual understanding" or text that encompasses "language understanding[2]. Benchmarking, the practice of evaluating ML models against standardized datasets, has evolved from humble beginnings in the 1980s and 1990s to become the primary mechanism through which progress in the field is measured and compared[3]. As[2] note, the inherent assumption underlying these practices is that performance on these datasets represents generalizable capabilities rather than narrow adaptations to the specific characteristics of the benchmark itself. Data sovereignty, in contrast, emphasizes the right of communities-especially Indigenous and historically marginalized ones-to maintain authority over data related to their territories, knowledge systems and cultural practices[4]. The concept has gained particular traction within Indigenous communities globally, who have articulated principles for data governance that center collective benefit and self-determination[5]. Emerging from this framework, Decolonial AI represents a growing field of critical analysis that seeks to identify and dismantle colonial power structures embedded within artificial intelligence technologies[6]. The problem at the heart of this inquiry is that most datasets portrayed as universal reflect distinctly Western cultural norms, linguistic patterns and geopolitical priorities[3]. Image Net, for instance, contains problematic categorizations of people that reflect racial and gender stereotypes[2]. The GLUE benchmark for natural language understanding is overwhelmingly centered on English, with constructions that privilege particular linguistic and cultural backgrounds[7]. Even COCO, designed for object recognition, displays strong geographic biases toward Western urban environments[8]. These biases are not merely technical oversights but have profound real-world impacts. They contribute to the erasure of cultural context and diverse knowledge systems, as local and Indigenous ways of categorizing and understanding the world are subsumed under supposedly universal taxonomies[9]. This erasure often results in direct harm to marginalized communities through misrepresentation or invisibility[10]. Moreover, the practice of data extraction from diverse communities without appropriate governance structures legitimizes extractive practices that echo colonial resource appropriation[6]. Perhaps most fundamentally, these practices perpetuate what Miranda Fricker (2007) terms epistemic injustice-the devaluing of certain groups' knowledge and interpretive resources. The purpose of this paper is to critique the myth of universality in benchmark datasets and advocate for context-sensitive alternatives that respect data sovereignty and epistemic diversity. I argue that the concept of universality in datasets not only misrepresents the inherently situated nature of knowledge but also reinforces structural inequities in global AI development. By applying decolonial frameworks to the analysis of benchmark datasets, this paper aims to reveal the epistemological assumptions that underpin claims of universality and to suggest more equitable alternatives.

**The Paper Proceeds as Follows:** Section 2 outlines the methodological approach, drawing on decolonial theory and data sovereignty frameworks. Section 3 presents a critical analysis of benchmark datasets, examining their historical origins, construction politics, universality claims and relationship to data sovereignty. Section 4 synthesizes these findings into a broader argument for pluralistic data practices and proposes concrete alternatives. Section 5 concludes with a summary of key points and a call to action for researchers, institutions and funders.

## MATERIALS AND METHODS

This paper employs a critical literature review and conceptual analysis methodology to examine the relationship between universal benchmark datasets, data sovereignty and decolonial theory. Rather than presenting new empirical findings, this approach synthesizes existing scholarship to reveal underlying assumptions, power dynamics and epistemological frameworks that shape the creation and use of datasets in machine learning. This methodological choice reflects the need for theoretical groundwork in addressing questions of justice and sovereignty in AI development. The analysis is guided by four interrelated theoretical frameworks from decolonial and critical theory. First, Walter Mignolo's (2009) concept of "epistemic disobedience" provides a lens through which to examine how knowledge systems become hierarchically organized and how resistance to

dominant epistemologies can create space for alternative ways of knowing. Mignolo argues for "delinking" from Western epistemology's claims to universality, highlighting how knowledge is always situated within specific geo-historical contexts. This framework helps us understand how benchmark datasets, despite claims to universality, embed particular world views and marginalize others. Second[11] epistemologies of the South emphasizes the systematic suppression of knowledge produced outside Euro- American centers of power-what he terms epistemicide. De Sousa Santos argues for cognitive justice that recognizes the plurality of knowledge systems and resists the reduction of diverse epistemologies to Western scientific frameworks. This perspective illuminates how benchmark datasets often perform epistemicide by codifying Western categorizations and ignoring alternative ontological frameworks. Third, Miranda Fricker's (2007) work on epistemic injustice offers analytical tools to identify how power imbalances create unfair disadvantages in both the production and interpretation of knowledge. Her concepts of testimonial injustice (the devaluing of someone's word based on identity prejudice) and hermeneutical injustice (the lack of collective interpretive resources to make sense of experiences) help us understand how benchmark datasets can perpetuate injustice by excluding certain voices and interpretive frameworks. Finally, Philip Alston's work on digital colonialism, particularly his critiques as UN Special Rapporteur on extreme poverty and human rights, highlights how technologies can reinforce existing power hierarchies and create new forms of dependency and exploitation[12]. This framework helps connect the technical aspects of dataset construction to broader socio-political dynamics of global inequality. The literature included in this review spans several disciplines and was selected based on specific criteria. First, I examined seminal works on benchmark datasets in machine learning, focusing on papers that introduce major benchmarks (e.g., ImageNet, GLUE) as well as technical analyses of their properties. Second, I included critical analyses of fairness and bias in machine learning, particularly those that go beyond technical fixes to examine structural and epistemic issues. Third, I incorporated literature on Indigenous data sovereignty, particularly declarations, principles, and frameworks developed by Indigenous scholars and communities. Finally, I drew on decolonial and critical theory literature to provide conceptual tools for analysis. This non-empirical approach is justified by several factors. First, questions of justice, sovereignty, and epistemic power cannot be adequately addressed through purely empirical methods., they require conceptual analysis that examines underlying assumptions and power dynamics. Second, the field needs theoretical foundations to guide empirical investigations, ensuring that technical solutions do not inadvertently reproduce problematic assumptions. Finally, this approach allows for a synthesis of insights across disciplines that might otherwise remain siloed. While empirical studies measuring specific biases in datasets are valuable, this paper takes a step back to examine the broader epistemological frameworks that shape how datasets are conceived, constructed and deployed. By bringing together technical literature on benchmark datasets with critical theory and Indigenous data governance frameworks, this analysis aims to provide a more comprehensive understanding of how data practices in machine learning relate to questions of sovereignty, justice and decolonization.

## RESULTS AND DISCUSSIONS

**Historical Origins of Benchmark Datasets:** The development of benchmark datasets in machine learning reveals a trajectory that has increasingly centralized certain institutions, epistemologies and geographic regions. Understanding this history is essential to recognizing how seemingly technical decisions reflect broader power dynamics in knowledge production. ImageNet, one of the most influential benchmark datasets, emerged from Stanford University in 2009, created by Fei-Fei Li and colleagues with the goal of advancing visual recognition systems through large-scale data[13]. Its taxonomy was derived from WordNet, a lexical database developed at Princeton University that organizes English words into hierarchical categories. This origin already embeds a particular linguistic and cultural framework-namely, Anglo-American categorization systems-into what would become a global standard for computer vision[14]. The General Language Understanding Evaluation (GLUE) benchmark, similarly, was developed primarily by researchers at New York University, the University of Washington and DeepMind in 2018[7]. It consolidates nine English-language natural language understanding tasks, positioning English linguistic patterns and cultural references as the default standard for evaluating language models. As noted by[15], this English-centricity is rarely acknowledged as a limitation in technical papers, instead being treated as an unmarked universal standard. These datasets were not created in a vacuum but within specific institutional contexts and funding structures. The development of ImageNet was supported by major tech companies and U.S. government agencies, including DARPA. GLUE's development similarly involved corporate interests through Deep Mind (Alphabet/Google). This concentration of resources shapes not only what gets built but also what questions are deemed worthy of investigation. As[3] note, The values embedded in these systems reflect the perspectives, priorities and prejudices of a narrow slice of humanity: their creators

(p. 2). A critical examination of these origins reveals implicit assumptions about universality and objectivity. The creators of these datasets often position them as neutral platforms for evaluation rather than as culturally situated artifacts. ImageNet's creators described it as a database of object classes that aims to populate the majority of the 80,000 synsets of WordNet with an average of 500-1000 clean and full resolution images[13], presenting it as a comprehensive visual taxonomy without acknowledging its cultural specificity. This positioning as universal standards rather than culturally situated artifacts has profound implications. It naturalizes particular ways of categorizing the world, making alternatives invisible or marking them as deviations from an unmarked norm. As[9] argues, this is characteristic of colonial knowledge systems, which present themselves as universal while concealing their geopolitical locations and interests.

**The Politics of Dataset Construction:** The process of constructing datasets-defining categories, collecting examples and establishing evaluation metrics-is deeply political, embedding particular values and perspectives that shape what machine learning systems learn and how they are evaluated. Category definition processes in benchmark datasets often reflect cultural assumptions about what constitutes meaningful divisions of the world[2]. Documented how ImageNet's person subtree included problematic categorizations based on appearance, character judgments and socioeconomic status-categories that reflected social biases rather than visual distinctiveness. Their analysis revealed 1,593 problematic synsets (out of 2,832) in the person categories that were potentially offensive, subjective, or reflected harmful stereotypes. These problematic categorizations are not merely theoretical concerns but have real impacts[14]. Excavating AI project documented how ImageNet categorized people with labels like bad person, closet queen, escort, or kleptomaniac, often based on appearance alone. Such categorizations not only misrepresent individuals but also encode harmful stereotypes that machine learning systems then reproduce and amplify. Linguistic biases are similarly embedded in natural language processing datasets. GLUE's focus on English text ignores the rich linguistic diversity of human communication, implicitly positioning English linguistic patterns as the standard for language understanding[15]. Moreover, even within English, these datasets often reflect the linguistic patterns of specific communities-typically educated, urban populations-rather than the full diversity of the language[16]. Geographic and demographic skews in visual datasets further demonstrate how seemingly neutral data collection processes reproduce existing inequalities[10]. Analyzed how biodiversity is misrepresented in ImageNet, finding significant biases toward North American and European flora and fauna, and noting that ImageNet-1k presents significant geographical and cultural biases, both in the labels used for non-human animal classes and in the images included in these classes (p. 5). Similarly[17], found that image datasets collected through internet searches disproportionately represent wealthy regions and populations. These biases are further compounded by the labor practices involved in dataset construction. Many benchmark datasets rely on crowd workers-often from the Global South-to label and categorize data according to taxonomies and guidelines developed in Western research institutions[18]. This creates a situation where those contributing their labor to datasets have little input into how categories are defined or how their work will ultimately be used. The politics of dataset construction thus reveals how power operates through seemingly technical decisions about categorization, collection and evaluation. Far from being neutral technical artifacts, benchmark datasets encode particular ways of seeing and knowing that reflect the social contexts in which they were created. As[6] argue, these practices often reinscribe colonial power dynamics by extracting data from diverse contexts while centralizing control over how that data is organized and interpreted.

**The Illusion of Universality:** The claim that benchmark datasets represent universal standards for evaluating machine learning capabilities rests on epistemological assumptions that deserve critical scrutiny. This section examines how the concept of universality in datasets misrepresents the inherently situated nature of knowledge and reproduces colonial epistemic hierarchies. A theoretical critique of universality claims begins with recognizing that all knowledge is situated-produced from particular locations, perspectives and histories rather than a god's eye view that transcends context[19]. As[9] argues, claims to universality often mask the particularity of Western epistemology, presenting it as an unmarked standard rather than one knowledge system among many. Applied to benchmark datasets, this insight reveals how claims to measure general language understanding or visual recognition conceal the specific cultural and linguistic frameworks within which these concepts are operationalized. Case studies of contextualized failures provide concrete evidence of this situatedness. For example[20] demonstrated how image recognition systems trained on ImageNet performed significantly worse on images from lower-income countries, showing that supposed general visual recognition capabilities were in fact specific to particular cultural and geographic contexts. Similarly[21], Found that object recognition models perform worse on images from Global South contexts, where everyday objects and scenes differ from those predominant in Western-centric training data. These failures reflect the marginalization of non-Western

knowledge systems in the construction of benchmark datasets. Traditional taxonomies, categorization systems, and epistemologies from non-Western contexts are rarely incorporated into benchmark design, despite their potential to offer valuable alternative frameworks[22]. This exclusion reproduces what de[11] terms the abyssal line of modern thinking-a division that renders certain forms of knowledge visible and credible while making others invisible or dismissing them as superstition or folklore. The replication of colonial power dynamics in data extraction further undermines claims to universality. As[23] argue, contemporary data practices often constitute a form of data colonialism, extracting value from human experience without meaningful consent or benefit-sharing. Benchmark datasets frequently rely on data harvested from diverse global contexts, yet the power to define categories, determine uses, and benefit from insights remains concentrated in Western institutions[6]. This asymmetry echoes colonial patterns of resource extraction and knowledge appropriation, where raw materials (data) are extracted from peripheral regions to generate value in core economies. The illusion of universality thus serves to naturalize particular knowledge systems while rendering alternatives invisible or subordinate. By presenting Western epistemological frameworks as unmarked standards rather than culturally specific approaches, benchmark datasets contribute to[24] terms hermeneutical injustice-the lack of collective interpretive resources to make sense of experiences. This has profound implications not only for the performance of ML systems across diverse contexts but also for which problems are deemed worthy of attention and which solutions are considered valid.

**Data Sovereignty and Community Rights:** In response to the problematic assumptions and practices embedded in universal benchmark datasets, alternative frameworks centered on data sovereignty and community rights have emerged. These approaches offer not only critiques but constructive alternatives for more ethical and inclusive data practices. Indigenous data governance models provide particularly rich frameworks for reconceptualizing the relationship between communities and data. The CARE Principles for Indigenous Data Governance-Collective Benefit, Authority to Control, Responsibility and Ethics-offer a people-centered approach that complements the technically-oriented FAIR principles (Findable, Accessible, Interoperable, Reusable)[5]. While FAIR principles focus on data management, CARE principles emphasize who benefits from data, who controls data governance and how values guide data use. As[5] explain, The CARE Principles address important considerations for modern data ecosystems and across data lifecycles that support both innovation and Indigenous self-determination (p. 3). These principles find practical expression in community-led dataset initiatives that prioritize local control and benefit. The Masakhane project for African languages exemplifies this approach, bringing together researchers across Africa to develop NLP resources that reflect the linguistic diversity of the continent rather than imposing external standards[25]. As stated on their website, Masakhane is guided by principles including African-centricity, Ownership and Data sovereignty, emphasizing that Africans should be able to decide what data represents our communities globally, retain ultimate ownership of that data and know how it is used. Similarly, Mozilla Common Voice has developed a participatory approach to creating multilingual voice datasets, working directly with language communities to ensure their voices are represented on their own terms. As they explain, Most voice datasets are owned by companies, which stifles innovation. Common Voice is the most diverse open voice dataset in the world. This approach reframes dataset creation as a collaborative process centered on community needs rather than extractive collection of training data. Decolonial approaches to data stewardship extend beyond specific projects to reimagine the relationships and values that guide data practices. Drawing on Mignolo's concept of "delinking," these approaches seek to create space for epistemological diversity by challenging the hegemony of Western data frameworks[6]. This involves not only technical changes but transformations in governance structures, decision-making processes and underlying values. The relationship between data sovereignty and cultural preservation highlights how data practices are inseparable from broader questions of cultural autonomy and self-determination. As Indigenous scholars have argued, data sovereignty is an expression of broader sovereignty rights-the ability of communities to govern themselves according to their own values and priorities[4]. When communities control how they are represented in datasets, they can ensure that these representations reflect their self-understanding rather than external stereotypes or classifications. These alternative approaches demonstrate that more ethical and inclusive data practices are not only possible but already emerging through community-led initiatives. By centering principles of sovereignty, community control and epistemological diversity, these approaches challenge the assumption that datasets must be universal to be valuable. Instead, they suggest that the greatest value may come from datasets that explicitly acknowledge their situatedness within particular cultural contexts and governance frameworks.

**Synthesis of Key Argument:** The preceding analysis reveals how the concept of the universal dataset

functions as a myth that justifies epistemic colonization in machine learning. By presenting culturally specific categorizations, linguistic patterns and visual representations as unmarked standards, benchmark datasets naturalize particular ways of knowing while marginalizing others. This process echoes[9] describes as the colonial matrix of power, where knowledge production becomes a site of domination that extends beyond formal colonial structures. The myth of universality serves several functions within machine learning research. First, it simplifies evaluation by implying that performance on a single benchmark meaningfully represents capabilities across diverse contexts. Second, it enables comparisons between models without addressing the complex question of what constitutes meaningful performance in different cultural and linguistic settings. Finally, it obscures the political nature of dataset construction, presenting what are fundamentally value-laden decisions as technical necessities. This analysis suggests that advancing machine learning requires balancing technical performance with cultural legitimacy and context-sensitivity. As[16] argue, Technical solutions alone cannot address issues rooted in social inequalities (p. 7). Rather than pursuing ever-higher accuracy on existing benchmarks, researchers might instead focus on developing systems that perform well across diverse contexts and respect the epistemic frameworks of the communities they serve. The implications for the field of ML research and applications are profound. If benchmark datasets embed particular cultural perspectives rather than universal standards, then progress in the field cannot be measured solely by improvements on these benchmarks. Instead, evaluating progress requires attending to how well systems serve diverse communities, respect varied epistemological frameworks and distribute benefits equitably across global contexts. This represents a fundamental shift from pursuing performance on standardized tests to pursuing justice and inclusivity across applications.

**Proposed Alternatives:** In response to these critiques, several alternative approaches offer promising directions for more ethical and inclusive data practices in machine learning. The concept of data pluriverses proposes multiple coexisting dataset paradigms rooted in specific cultural contexts, challenging the assumption that a single dataset can meaningfully represent a domain. Drawing on notion of the pluriverse as a world where many worlds fit, this approach recognizes that different communities may categorize, interpret and value data in fundamentally different ways. Rather than seeking to harmonize these differences into a universal standard, data pluriverses would maintain the integrity of distinct epistemological frameworks while facilitating communication between them. Co-design frameworks for local benchmarking offer practical methodologies for creating datasets that reflect community priorities and epistemologies[25]. Describe how the Masakhane project involves local communities not just as data sources but as full participants in defining research questions, designing datasets and interpreting results. This approach ensures that benchmarks reflect the linguistic patterns and cultural contexts they purport to measure, rather than imposing external standards. Decentralized dataset stewardship by affected communities represents a governance approach that aligns with principles of data sovereignty. Rather than centralizing control over datasets in research institutions or corporations, this model distributes authority to the communities represented in the data. As[5] argue, Authority to Control affirms the rights and interests of Indigenous Peoples in governing the collection, access and use of their data (p. 3). This approach challenges dominant models of data ownership and control, positioning communities as rights-holders rather than merely stakeholders. Integration of CARE principles into mainstream ML practices offers a framework for operationalizing these alternatives. The principles of Collective Benefit, Authority to Control, Responsibility and Ethics provide concrete guidance for dataset creation, governance, and use that respects community rights and promotes equitable outcomes. Importantly, these principles are complementary to technical best practices like the FAIR principles, suggesting that technical excellence and ethical governance can and should go hand in hand. These alternatives share a commitment to pluralism and locality over universalism and centralization. They recognize that meaningful progress in machine learning requires not just technical sophistication but ethical and epistemological diversity. By embracing multiple ways of knowing and distributing power across communities, these approaches offer a vision of AI development that supports rather than undermines data sovereignty.

**Future Implications:** The shift from universal to pluralistic approaches to datasets has significant implications for the future of AI development and governance. For AI justice and accountability, the recognition that datasets embed particular cultural perspectives challenges simplistic approaches to fairness. Rather than merely balancing representation within existing categories, justice requires questioning the categories themselves and the power dynamics that produce them. Accountability, likewise, must extend beyond technical metrics to consider how systems affect communities' ability to define themselves and determine how they are represented. Localized governance structures become essential in this context, as they enable communities to exercise

meaningful control over how data about them is collected, interpreted and used. This may involve developing new institutions and practices that bridge technical expertise with community knowledge and priorities. The CARE principles offer guidance here, emphasizing that governance should promote collective benefit and respect community authority. For large language models (LLMs), which represent a significant frontier in contemporary AI development, these critiques take on particular urgency. Current LLMs are trained on massive datasets that reproduce and amplify existing biases in internet text, scholarly publications and other sources[1]. Applying decolonial perspectives to these models requires not just technical interventions but fundamental reconsiderations of what constitutes knowledge and who has the authority to define it. The future of benchmarking in a pluralistic data ecosystem may involve moving beyond the pursuit of universal standards to embrace contextual evaluation. As[1] suggest, this might include creating benchmarks that are explicitly contextualized, rather than purporting to measure general abilities (p. 615). It might also involve developing new metrics that value cultural relevance and community benefit alongside traditional technical measures. This approach has limitations that deserve acknowledgment. Pluralistic approaches may be more resource-intensive than centralized ones, requiring engagement with diverse communities and epistemologies. They may also complicate comparisons between models, making it more difficult to track progress in straightforward ways. However, these challenges reflect the inherent complexity of developing technologies that serve diverse human contexts rather than technical limitations to be overcome. Future research in this area might explore how pluralistic approaches can be implemented at scale, how community governance can be supported without imposing unreasonable burdens on already marginalized groups and how different epistemological frameworks might inform not just dataset creation but algorithm design. There is also important work to be done in developing metrics that can meaningfully assess how well systems respect data sovereignty and epistemological diversity.

## CONCLUSION

This paper has critically examined the concept of universal benchmark datasets in machine learning through the lens of data sovereignty and decolonial theory. The analysis has revealed how datasets positioned as universal standards often embed particular cultural perspectives, linguistic patterns and epistemological frameworks, marginalizing alternative ways of knowing and reproducing colonial power dynamics in data practices. The myth of universality in benchmark datasets is not merely a technical oversight but a reflection of deeper epistemic hierarchies that privilege certain forms of knowledge while rendering others invisible or subordinate. By presenting Western categorizations and frameworks as unmarked standards rather than culturally specific approaches, these datasets naturalize particular ways of seeing and knowing while excluding alternatives. The urgency of dismantling universalist assumptions in AI cannot be overstated. As machine learning systems increasingly shape critical aspects of social, economic and political life, the epistemological assumptions embedded in their training data have far-reaching consequences. The perpetuation of colonial knowledge hierarchies through supposedly universal datasets threatens to amplify existing inequalities and further marginalize communities that have historically been excluded from technology development. Researchers, institutions, and funders have vital roles to play in fostering more ethical and pluralistic data practices. Researchers can critically examine the epistemological assumptions in their work, engage meaningfully with diverse communities and develop methods that respect epistemological diversity. Institutions can implement governance structures that distribute power and benefits equitably, ensuring that communities have meaningful control over how they are represented in datasets. Funders can prioritize projects that center data sovereignty and community governance, creating incentives for more ethical and inclusive approaches. The vision for more ethical and pluralistic data practices goes beyond technical fixes to reimagine the relationships between technology development and the communities it affects. By embracing data pluriverses, co-design methodologies, decentralized stewardship and principles like CARE, the field can move toward approaches that respect diverse epistemologies and support community self-determination. In the words of de Sousa Santos (2014), Another knowledge is possible (p. ix). By recognizing the limitations of universal datasets and embracing pluralistic alternatives, machine learning can contribute to a more just and inclusive technological future-one where many ways of knowing are valued and where communities have the authority to determine how they are represented in the datasets that increasingly shape our world.

## REFERENCES

1. Bender E.M., T. Gebru, A. McMillan-Major and S. Shmitchell., 2021. On the Dangers of Stochastic Parrots: Can language models be too big? Proc. 2021 ACM Conf. Fairness, Accountability, Transparency, Vol. 10.1145/3442188.3445922.
2. Yang K., K. Qinami, L. Fei-Fei, J. Deng and O. Russakovsky., 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. Proc. 2020 Conf. Fairness, Accountability, Transparency, Vol. 10.1145/3351095.3375709.

3. Birhane A., P. Kalluri, D. Card, W. Agnew, R. Dotan and M. Bao., 2022. The Values Encoded in Machine Learning Research. 2022 ACM Conf. Fairness Accountability Transparency, Vol. 10.1145/3531146.3533083.
4. Kukutai T. and J. Taylor (Eds.)., 2016. Indigenous data sovereignty: Toward an agenda. ANU Press., Vol. 0. 10.22459/CAEPR38.11.2016.
5. Carroll S.R., I. Garba, O.L. Figueroa-Rodríguez, J. Holbrook and R. Lovett *et al.,* 2020. The CARE Principles for Indigenous Data Governance. Data Sci. J., Vol. 19. 10.5334/dsj-2020-043.
6. Muldoon J. and B.A. Wu., 2023. Artificial Intelligence in the Colonial Matrix of Power. Philosophy and Technol., Vol. 36: 10.1007/s13347-023-00687-8.
7. Wang A., A. Singh, J. Michael, F. Hill, O. Levy, and S.R. Bowman., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proc. Wor. Bla. Anal. Inter. Neur. Network. NLP Vol. 10.48550/ arXiv.1804.07461.
8. Mokhtar O., N. Rokni and A. Sivertsen., 2022. The bias in the machine: Dataset biases in computer vision systems for autonomous vehicles. .Proc. Conf. Fairn. Acco. Transparency., 1233-1244.
9. Mignolo, W.D., 2009. Epistemic Disobedience, Independent Thought and Decolonial Freedom. Theory, Cult. and Soc., Vol. 26: 10.1177/ 0263276409349275.
10. Luccioni A.S. and D. Rolnick., 2023. Bugs in the Data: How ImageNet Misrepresents Biodiversity. Proc. AAAI Conf. Artif. Intell., Vol. 37: 10.1609/aaai.v37i12.26682.
11. Santos B.D.S., 2014. Epistemologies of the South: Justice against epistemicide. Routledge., Vol.
12. Alston P., 2019. Report of the Special Rapporteur on extreme poverty and human rights. United Nations General Assembly., Vol. 74.
13. Deng J., W. Dong, R. Socher, L.J Li, K. Li and L. Fei-Fei., 2009. ImageNet: A large-scale hierarchical image database. In: In 2009 IEEE conference on computer vision and pattern recognition., IEEE., 0 pp: 248-255.
14. Crawford K. and T. Paglen., 2019. Excavating AI: The politics of images in machine learning training sets. The AI Now Institute., Vol.
15. Bender E.M. and B. Friedman., 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Trans. Assoc. Comput. Linguistics, Vol. 6: 10.1162/tacl_a_00041.
16. Shah D.S., H.A. Schwartz and D. Hovy., 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. Proc. 58th Annual Meeting Assoc. Comput. Linguistics, Vol. 10.18653/v1/2020.acl-main.468.
17. De Vries T., I. Misra, C. Wang and L. van der Maaten., 2019. Does object recognition work for everyone? Proc. IEEE/CVF Conf. Comp. Vis. Pat. Recog. Workshops., Vol. 10.48550/arXiv. 1906.02659.
18. Gray M.L. and S. Suri., 2019. Ghost work: How to stop Silicon Valley from building a new global underclass. Houghton Mifflin Harcourt., Vol.
19. Haraway D., 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Stud., Vol. 14: 10.2307/3178066.
20. Shankar S., Y. Halpern, E. Breck, J. Atwood, J. Wilson and D. Sculley., 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world.
21. Ramakrishnan S., A. Agrawal and S. Lee., 2017. Addressing cross-country heterogeneity in computer vision systems with adversarial networks. Proc.SIG.Con. Com.Sust. Soci., 1-12.
22. Bidese H., B. Ásbjörnsdóttir and J. Qi., 2020. Kila Àrúgbó (and other stories): Knowledge organization and the politics of knowledge production. Journal of Critical Library and Information Studies., Vol. 3.
23. Couldry N. and U.A. Mejias., 2019. The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press. 10.1515/9781503609754.
24. Fricker M., 2007. Epistemic injustice: Power and the ethics of knowing. Oxford University Press., Vol.
25. Nekoto W., V. Marivate, T. Matsila, T. Fasubaa and T. Fagbohungbe et al., 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. Findings Assoc. Comput. Linguistics: EMNLP 2020, Vol. 2020: 10.18653/v1/2020.findings-emnlp.195.