A Capabilities Approach to Fairness and Inclusion in Language Technologies

HELLINA HAILU NIGATU, University of California at Berkeley, USA
ZEERAK TALAT, University of Edinburgh, UK

Mainstream Natural Language Processing (NLP) has largely ignored the majority of the world's languages. However, the increasing sophistication of English language technologies is directing efforts towards the disregarded masses of languages. When porting language technologies to new languages, researchers also have to contend with the risk of exposing language users to risks imposed by the technology. While there are existing safeguards and mitigation strategies for English, these measures may not port well to understudied languages due to cultural shifts or technical constraints. Further, by porting English language datasets for measurement and mitigation, we disregard (1) the community needs for language technologies; and (2) the harms arising from bias and fairness issues within the context of a given community. In this paper, we seek to add nuance to the discussion around access to language technology—and more widely, machine learning—fairness, inclusion, and bias through the lens of the Capabilities Approach. The Capabilities Approach emphasizes *what people are capable of achieving*, given their intersectional, social, political, and economic contexts. Thus, through its emphasis on the ability to realize use, it offers a contrast to asking merely *what resources are (theoretically) available to a community*. We show how adopting this framework brings forth the intersectional identities of users of language technologies through three case studies, and argue that by using the capabilities approach, we can gain a more complete picture of how to ensure that machine learning and language technologies are accessible and useful to communities that have so far been neglected by the research community. We hope that our work inspires further attention to the nuances of social and cultural realities of diverse users in creating fair language technologies.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Natural language processing**; Language resources.

Additional Key Words and Phrases: low-resourced languages, language technologies, fairness, capabilities approach, NLP

ACM Reference Format:
Hellina Hailu Nigatu and Zeerak Talat. 2026. A Capabilities Approach to Fairness and Inclusion in Language Technologies. 1, 1 (January 2026), 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Research advances in Natural Language Processing (NLP) have improved performance for several language technologies used in our daily lives. However, despite its many advances, NLP research has continually left out a wide swath of languages and communities from its development [45, 67] and evaluation [66, 111].

According to Joshi et al. [45], over 88% of the world's languages are continually excluded from research and development in language technologies. Many of these languages are spoken by African, Asian, Indigenous, and Latin American communities, which constitute the vast majority of the world's population and languages —the African continent alone is home to over 2000 of the 7000 languages of the world [32]. As a result, speakers of these languages

Authors' Contact Information: Hellina Hailu Nigatu, hellina_nigatu@berkeley.edu, University of California at Berkeley, Berkeley, CA, USA; Zeerak Talat, University of Edinburgh, Edinburgh, UK, z@zeerak.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

are often left with no access to technology in their own language [88, 89, 99], or with technologies that are rendered non-functional by their poor performances [27]. Correspondingly, these languages and their speakers are also under-represented in evaluation studies [111]. Prior work shows that the majority of study findings published in popular HCI venues are based on WEIRD participants,[1] demonstrating a lack of representation of the majority of speakers of the world's languages in design and evaluation studies [55, 97].

Recent efforts have been targeted towards increasing the representation of Global Majority communities in the development and evaluation of language technologies, including by building multilingual models [e.g. 28, 30, 46], collecting data in under-represented languages [e.g. 35, 54], and evaluating the performance of State-Of-The-Art (SOTA) technology in several languages [e.g. 8, 73]. However, aggressively including languages spoken by previously excluded and continually exploited communities in technologies that are primarily designed by and for Western languages extend colonial control and exploitation of the communities without proper engagement or substantial benefit to community members [12, 13]. Further, building and adopting language technology for understudied languages has several challenges. First, the socio-technical evaluation of language technologies–even if we only evaluate for a single, high-resourced language like English–is a difficult task that requires us to engage with several social axes [17, 101]. This is further complicated in Global Majority communities, which are diverse in languages, cultures, and socio-economic standings[10, 101]. Second, in moving from excluding the majority of the world's languages to blindly adopting what we make for English, we risk importing the same harms we have at best mitigated and at least measured for English [37, 63]. For instance, Yong et al. [114] showed how prompting GPT-4 in low-resource languages circumvents guardrails that are effective in English. Third, when our research agenda is shaped by the tools and resources that are available and working for better-studied languages like English, we risk ignoring the breadth of a community's needs for other and new language technologies [27]. These three challenges call into question business-as-usual for developing research artifacts for low-resource languages. We argue that it is necessary to take a holistic approach to developing and evaluating language technologies that make explicit the diverse identities and socio-political forces of the Global Majority in order to avoid building technologies that are impractical and unusable for Global Majority users.

In this paper, we view fairness, bias, and inclusion in language technologies through the lens of the Capabilities Approach (CA) [90, 95]. The Capabilities Approach is a framework from developmental economic studies that centers *what people are capable of achieving*, given their intersectional social, political, and economic contexts, instead of *what resources are (theoretically) available to them.* Adopting this framework would allow us to shift our question from "How does this language technology impact this community?" to "What language technology can this community make use of?" By doing so, we ensure our design and evaluation are situated within the context of the community.

Our paper makes three key contributions. First, we present the theoretical background for the Capabilities Approach and provide practical considerations for its applications to research on language technology (Section 3). Second, we discuss three case studies in Section 4 to illustrate how the CA affords an approach that is sensitive to diverse contexts and people. We use our case studies to outline the benefits of the capabilities approach, show its relationship to multilingual and multicultural evaluation, and demonstrate how it affords meaningful collaboration with community members in defining and measuring the fairness and inclusion in language technologies. Third, we build on our case studies and the theoretical background to synthesize a discussion on the benefits and limits of adopting the approach.

We argue that for research and development of language technologies to be useful to under-represented groups, a different approach must be taken than for high resource languages. It is especially pertinent that we question our

[1] WEIRD is an acronym for Western, Educated, Industrialized, Rich, and Democratic. The term has been used by prior work to demonstrate the hegemony in different fields [e.g. 40, 97]

current approach in building language technologies as it is more likely for high resource language speakers to develop and publish artifacts for low-resource scenarios, than for speakers of low resource languages to build the technologies for their own languages [39, 67]. The capabilities approach is a promising avenue for reconfiguring our ways of working with language communities, giving us an alternative to one-size-fits-all approaches that fail in the context of Global Majority communities.

2 Background

In this section, we provide background on communities of the Majority Worldand briefly describe the history of their exclusion from mainstream research and development.

2.1 Identities of the Majority World

The "Majority World" refers to the peoples of African, Indigenous, Asian, and Latin American descent. The term was coined by Shahidul Alam [7] to emphasize that these populations constitute the majority of the world's population but are continually referred to with marginalizing terminology[2] Under the umbrella term "Majority World" lie several cultures, languages, and communities, all with varying levels of economic, political, and social fabrics [48].

The majority of the world's languages–and their dialects–are spoken by communities from African, Asian, and Indigenous populations [70]. However, these communities have historically been and continue to face the brunt of (neo-)colonization [13]. Among its several impacts, colonization has resulted in the loss and damage of traditional knowledge preservation systems—including damage to language preservation [47, 56, 62]. Language policies from colonial times have shaped the formation of education policies and language preservation tactics, which in turn have put constraints to resources that are (digitally) available for the languages of the Majority [11, 59].

Communities of the Majority World face systemic barriers to access services such as healthcare and education [21, 24, 110]. These barriers are especially pronounced for those at the intersection of several marginalized social identities–for instance, barriers to access to healthcare for women in India are particularly pronounced for women from minoritized casts and lower-income backgrounds [80]. As the identities of the Majority World communities are diverse and, at times, distinct from those of the West, the axes along which disparity can occur also diverge from those popularly studied in Western contexts. Asiedu et al. [10] demonstrate that while "axes of disparity" such as race and ethnicity may apply globally, contextual axes like colonial history and country of origin also shape individuals' experiences for African communities. As a consequence, projects focusing on the Majority World may require different approaches than those for Western contexts.

2.2 Exclusion of the Majority World

The AI research landscape in general and NLP research in particular has increasingly adopted methods that rely on large amounts of datasets [46, 71]. Training such models requires computational resources that are not readily available outside of a small number of large institutions, which are mostly located in Western countries, e.g., large software companies such as Google and Meta, and certain Western universities [92]. As a result of this resource-hungry trend, most of the State-Of-The-Art models do not work for the majority of the world's languages [45]. Further, attempts to make them work often rely on machine-translated documents using translation models that are similarly data intensive [108]. To describe the resource disparity among languages, the NLP research community has adopted the

[2] For instance, terms like "Third World" that are based on economic hierarchies insinuate backwardness [48].

terminology “low-resource” vs “high-resource” [69]. Many of the Global Majority languages are categorized as low-resourced with data scarcity [e.g. 31, 33, 50], computational resources [e.g. 5, 91], and the number of speakers [e.g. 34, 79] attributed as some reasons for categorizing them as such.[3]

In the current research landscape, Majority World communities experience harm and marginalization through (1) State-Of-The-Art models and methods not accounting their languages [72], (2) direct exploitation of their labor for training and moderating models [64], and (3) harm caused by failures in existing models [114]. Additionally, Global Majority communities are also disproportionately affected by the environmental impacts of Artificial Intelligence development, through resource extraction and high energy consumption by the AI industry[87]. The history of mainstream academic research is also tainted by parachute studies—where non-Western researchers are not given their due credit in academic publications and research [4].

Some efforts to remedy the lack of representation of diverse languages in NLP are targeted towards collecting data in a particular language(s) [e.g. 9] or adopting existing (usually English) datasets through human or machine translation [e.g. 38]. However, data collection processes by themselves could lead to harm, with risks of minimal compensation and lack of ownership over the collected datasets resulting in exploitation of the communities [4, 52, 84]. While human translation can have similar issues with a lack of compensation or ownership for the language community, machine translation poses a particular risk: machine translated documents often contain mistranslations, inaccuracies, and “translationese” [106, 109] and thus introduce an additional risk of representing languages distinct from their real-world usage. Further, using translated English text does not guarantee that the content will be culturally or contextually relevant to the community [113].

As we embed language technologies into diverse contexts, we risk perpetuating and reinforcing harm and further exclusion, especially to those already marginalized [17, 57] Furthermore, our current evaluation schemes do not account for the several interacting identities of the Majority World [10, 101]. In moving towards an inclusive development and evaluation ecosystem, one important question to ask is: **how do we make the diverse and entangled identities and social fabrics of the ‘Majority World’ explicit in our design and evaluation?**

3 The Capabilities Approach

In this paper, we propose adopting the Capabilities Approach to how we design and evaluate language technologies for languages of the Majority World. The Capabilities Approach is a framework from developmental economic studies proposed by Amartya Sen [94, 96]. The framework posits that if the “freedom to achieve well-being is of primary moral importance”, then we should consider well-being in terms of what each individual is capable of utilizing [90]. The Capabilities Approach has two main axioms. The first axiom is that “freedom to achieve well-being is of primary moral importance,” i.e, it is first established by the Capabilities Approach that individuals’ freedom to achieve well-being is of central importance [90]. The second axiom is that “well-being”–and to what degree it can be achieved–depends on what an individual is capable of achieving under the social, cultural, political, and economic constraints they face [90]. Holding those two axioms together, the Capabilities Approach affords a lens through which we can understand well-being in relation to what is possible for each individual, conditioned on their socio-political contexts.

In this section, we will first provide two illustrative examples to solidify where and how we can apply the Capabilities Approach (Section 3.1). We will then use our second illustrative example to provide a set of definitions towards applying

[3] It is important to note that not all languages classified as low-resourced are spoken in Global Majority countries. For instance, Gaelic and Irish are studied as low-resourced languages[e.g 60] although they are spoken within Europe.

the Capabilities Approach in the context of language technologies in Section 3.2 followed by Section 3.4 where we show practical steps to applying the approach.

3.1 Illustrative Examples

In this section, we provide two illustrative examples to ground our discussion on how one can apply the Capabilities Approach to designing and evaluating language technologies. The Capabilities Approach asks us to first question: *what resources is the particular community/individual capable of utilizing?*

We first give a general example, inspired by prior work [77, 90, 95]. Consider a community where people have to walk far to access basic services, such as access to groceries, water, and pharmacies. Providing community members with bicycles could be one environmentally friendly solution to alleviate the problem of access. However, to identify whether our proposed solution can help individuals in the community achieve easier access, we first need to ensure that (1) people in the community can and want to ride bicycles, (2) the paths and infrastructures are suitable for riding bicycles, and (3) there is enough expertise to maintain the bicycles. Further, there may be community members who are not physically able to ride bicycles, and are thus left outside of the solution margins. With the Capabilities Approach, we first seek to make these assumptions explicit, allowing us to understand where and how our resources will fail to meet the community's needs, which members of the community that our proposal excludes, and why they are not able to benefit from the proposed solution. Core to the Capabilities Approach is understanding that what resources a person owns or can (theoretically) use is insufficient to assess what individuals can achieve through the resources that are available to them.

With the above illustrative example in mind, let us map a similar scenario where the resource we are trying to distribute is a particular language technology. Consider a community with limited access to healthcare information. One potential solution to enable further access to healthcare information could be to provide them with a chatbot that is tailored towards answering health-related queries. We first need to collect enough data to build a chatbot for this community. However, access to datasets, especially those in domains like the healthcare sector, is extremely limited for languages of the Global Majority [65, 74]. One common approach in the NLP research community is to translate datasets from English [6, 38]. Even if we assume the translation was done without error, there are several socio-economic constraints to consider for the chatbot to be useful for this community.

We first have to ask who has access to digital devices and internet access to make use of the chatbot. Prior work shows that, especially within Majority World communities, women and members of marginalized ethnic groups are less likely to have access to digital devices [58, 83], and would therefore have no or minimal benefits from a chatbot. Second, we need to consider what risk mitigation strategies exist for when the chatbot makes an error, e.g., presenting inaccurate information [25] or outright fabricated content [18]. Indeed, prior work shows that errors in language technologies (across English and other languages) may lead to clinical harm to patients when deployed in the healthcare sector, with Machine Translated systems resulting in mistranslations in doctor-patient communication[49] and speech recognition models used for transcribing doctor notes outputting fabricated content [20]. Deploying such technology in places where there is limited access to actual healthcare services might increase risk when there are critical errors in the output of the language technology [61]. Third, we should consider whether there are cultural taboos that would affect how users interpret health-related answers, especially if we are using translation to build our dataset. For instance, in some cultures, discussion of psychological distress could be considered taboo, requiring a special handling of language when communicating about such issues in the languages spoken by that community[105]. Prior failed technosolutionist

Term	Definition	Example
Capability	Doings and beings a person can achieve if they want to	access to a chatbot to answer health-related questions.
Capability Set	A set of capabilities a person can choose from	access to a medical chatbot, access to healthcare services
Functioning	Capabilities that have been realized	using a medical chatbot
Real Freedom	Having all the necessary means to achieve a capability	social standing not restricting access to resources
Conversion Factor	The resources that a person has to transform a capability into a functioning	digital device, internet connection, electric power, risk mitigation resources

Table 1. Definitions of terminology in the Capabilities Approach.

endeavors also lead us to question how reliable this solution will be in the long run, if the community members do not have the necessary means to maintain it [76].

3.2 Definitions

In this section, we provide definitions of terminology that would be necessary to map the Capabilities Approach to a new context. We rely on our illustrative example from the previous section (Section 3.1) to ground definitions of the Capabilities approach and terminologies associated with it.

3.3 Defining the Capabilities Approach

It is first important to establish that the Capabilities Approach is described as a framework–as opposed to a theory–as it is an approach that is deliberately under-specified to allow it to be adaptable to answering normative questions about well-being [90]. Starting with the name of the framework, *capabilities* refer to "doings and beings that people can achieve if they choose" [90]. Note that the approach relies on an individual's desire–i.e., a person can choose whether they want to achieve a capability. Once a capability becomes realized, it is called a *functioning*. Drawing from our example in Section 3.1, the opportunity–i.e., access and desire–to use the chatbot for a health query is a capability, while using the chatbot is a functioning. The set of capabilities an individual can choose from is called a *capability set.* In our example, a person's capability set may include access to the chatbot and access to healthcare services. For a person to realize a capability into a functioning, they must have *real freedom*—i.e., access to all the necessary means to achieve a capability. For instance, if a person wants to achieve the functioning of using the chatbot to answer a health-related query, they must have the freedom to use a digital device with proper power and internet connectivity. *Conversion factor* refers to the factors that influence the degree to which an individual can convert a capability to a functioning. For example, to achieve the functioning using the chatbot for a health query, a person's financial resources and social responsibilities can be considered conversion factors. In Table 1, we summarize the definition for the different terminologies that are needed to apply the capabilities approach in the context of language technologies.

3.4 Applying the Capabilities Approach

The Capabilities Approach has been applied to various fields of study that deal with resource allocation and justice. For example, scholars have used the framework to argue for environmental and climate justice [e.g. 85, 93], the right to education [e.g. 104], and access to healthcare [e.g. 107]. The Capabilities Approach has also been adopted to discussions around technological designsuch as in conversations around design policy [29] and design features of technological artifacts [75].

While adopting technological solutions is viewed by some as a means to extend one's capabilities'[53], what an individual is capable of achieving depends on the relative position they hold within their social context [76]. As a result, one set of solutions may empower some while ignoring–or even actively harming–others. As Oosterlaken [76] argues, applying the Capabilities Approach to how we design technology would call for methodological individualism [100], putting our focus on the individual circumstances that may lead to differing experiences. Adopting a Capabilities Approach lens to how we study technology would shift our focus to understand where and how our proposed intervention might disempower or further marginalize certain groups of society [76]. For example, if the medical health chatbot required identification documents or was subscription-based, it would exclude those without such documents or means to cover such costs–communities who often exist at the intersection of other forms of exclusion and marginalisation. Further, the Capabilities Approach also challenges the notion that it is possible for one set of solutions to work in all contexts [53, 76, 100]. Oosterlaken [76] particularly calls out the many failed cases of technological solutions exported to Majority World countries.

As we highlight in Section 2, one of the major challenges in studying and adopting technological solutions to Majority World contexts is that our current evaluation schemes do not account for diversity within or between communities. The Capabilities Approach offers a solution to this challenge as it centers human diversity, by acknowledging its existence and in questioning how diversity affects what conversion factors are available for an individual [75, 100]. When evaluating a technological solution, the Capabilities Approach implores us to understand–based on the social structures of the community–whose capability the technology will enhance and who it will exclude or harm.

The Capabilities Approach thus provides a shift from "'What resources have we allocated and how can we measure their impact?" to "What resources do community members have the capability to utilize?" thereby centering the needs of the community in question. Additionally, the very definition of the approach rests on individuals having real freedoms and choices (Section 3.2). Hence, by applying the Capabilities Approach, we can conduct evaluations that center the agency and the needs of community members rather treat communities as monolithic in their need for technologies. Below, we provide some guiding questions to consider when applying the Capabilities Approach to a specific community, particularly looking at language technology as a resource:

Q 1 What potential functioning does this resource (language technology) seek to bring to the community?

- **Q 1.1** In which way(s) will the potential functioning(s) enabled by the resource benefit individuals and the community as a whole?
- **Q 1.2** In which way(s) will potential functioning(s) enabled by the resource harm individuals or the community as a whole?
- **Q 1.3** In which way(s) will potential functioning(s) enabled by the resource shift power dynamics within the community?

Q 2 Does this community have the capability to use this resource (language technology)?

- **Q 2.2** What limits the community from choosing to realize the capability to a functioning?
- **Q 2.3** What conversion factors do the community members have at their disposal?

Q 2.4 What is the list of capabilities (both individual and community level) that the community members have in realizing these resources into a functioning?

Q 3 What other resources (language technologies) does the community have the capability to utilize?

Q 3.3 What are the community needs (for language technologies)?

Q 3.4 Which conversion factors have aided or prevented the community in turning capabilities into functioning for other resources?

Q 3.5 What forms of harm is the community dealing with that we are not measuring?

Q 3.6 Which axes of marginalization impact or split the community in the ability to turn capabilities into functionings?

Using the above guiding questions as a starting point, we can design evaluation methods that center meaningful community engagement and explicate the diverse social identities of individuals.

4 Case Studies

In this section, we provide three case studies to demonstrate how applying the Capabilities Approach to evaluate language technologies foregrounds community needs. When applying the CA, we must first identify the different stakeholders, i.e., members or groups within a community; then co-create their Capability Set(s), and finally list the Conversion Factors for each stakeholder. We provide two primarily analytical and one practical case studies to illustrate how the CA can be applied to NLP projects. First, in Section 4.1, we view a data collection and preparation effort for building speech technologies for the indigenous languages of Kanien'kéha through the CA lens. Then, in Section 4.2, we examine the Te Hiku NLP project, in which they have built language technologies for the Māori on the pillars of data sovereignty and direct community benefit. Finally, in Section 4.3, we show how the Capabilities Approach can afford insights into how social structures influence the efficacy of our interventions through a case study on content moderation in low-resourced languages.

4.1 Case 1: Respecting Community Values in Data Collection and Processing

In our first case study, we demonstrate how we can apply the CA to the data collection and planning stage of building language technologies explicate community values. We use a speech technology development project for Kanien'kéha, illustrating how community values were respected during data collection and preparation. Particularity, we use the work of Pine et al. [79] who built speech synthesis models for four indigenous languages including Kanien'kéha, as a step towards language revitalization.

Identifying the community of study. Kanien'kéha is the language of the Kanien'kehá:ka, an indigenous population that originally inhibited the Mohawk River Valley in present day New York and had since been displaced and scattered to present day Canada and parts of New York as a result of colonial imposition [36]. Their language is written with the Latin script, with an alphabet that was standardized in 1993 [36]. According to Ethnolouge, Kanien'kéha is labeled as endangered in terms of vitality, indicating a low number of speakers and as ascending in terms of digital language support, indicating an increase in the amount of digitally available resources for the language [2].

Listing the Capability Set for each stakeholder. There are some efforts to preserve the language and facilitate intergenerational transfer: for instance, there are community based organizations including the Onkwawenna Kentyohkwa adult immersion school that, since 1999, has been operating with "the goal of creating fluent speakers of Kanien'kéha (the "Mohawk" language)" [3]. Despite its success in teaching Kanien'kéha and serving as a blueprint for four other indigenous languages, the Onkwawenna Kentyohkwa adult immersion school does not receive continued government funding. Instead, the school relies on community funds, which can only cover basic operations [3]. In terms of digitally available data, there exists a translation of the bible and audio recordings of the translations that were compiled by five Kanien'kehá:ka translators [79]. Parts of this data were used in a project that sought to revitalize the language through speech synthesis [79].

Listing the Conversion Factors for each stakeholder. With a speaker of the language on the team of researchers and three additional language speakers, Pine et al. [79] aligned separate speech and text data of bible narrations by five Kanien'kehá:ka speakers using automatic alignment tools. While the data included 24 hours of audio, four of the five speakers had passed away, leaving only one speaker's data [79]. As per the custom of the Kanien'kehá:ka, the data from the speakers that passed away was not used for training the speech synthesis model. Additional data cleaning to remove code-mixed words due to the lack of digital tools to process code-switched data resulted in only 3.46 hours of speech data that was used in the project [79]. As the researchers point out, the state of the art speech synthesis models have large data size requirements, inflating the amount of data needed to build effective speech synthesis models [79]. Relying on methods that had less data requirements, Pine et al. [79] illustrated that they could achieve similar performance for speech synthesis with a data size difference of 10-fold. In approaching the development of language technologies with care, the researchers managed to develop a speech synthesis model which was in accordance with the customs of the Kanien'kehá:ka while addressing a need for language technologies to aid with language revitalization.

Analysis using the Capabilities Approach. Looking at the project by Pine et al. [79] through a CA lens, we observe the role of community agency in the planning and processing stage of building the language technology. Even when NLP papers include under-studied languages, they rarely engage speakers of these languages [39]. As a result, community values are not reflected in the planning and implementation of such projects [12]. As we have seen above, Pine et al. [79] offers an alternative to this approach by 1) having a research team that comprises of a speaker of the language, 2) including speakers in the data processing stage, and 3) respecting community values despite losing over 80% of the data. Further, the project illustrated how instead of finding ways to increase the data to match the requirement of the State-of-the-Art (SOTA) tool, using an approach that better matched the existing data constraints resulted in better, more efficient solutions. Through respecting community values, this project is directly in line with the CA, in that the planning stage of the project is building from what is already available to the community instead of attempting to find ways to match the requirements of the SOTA technologies.

4.2 Case 2: Building Sovereign Speech Technologies

For our second case study, we will show how a community-centered approach to research results in solutions that are aligned with answering the questions posed by the Capabilities Approach. Particularly, we will examine how Te Hiku Media collected data for the Māori language, Te Reo Māori, to engage in language restoration efforts while safeguarding their data from being co-opted by Big Tech. We choose this case, as it has been highlighted as a rare example of decolonial machine learning by Birhane and Talat [15].

Identifying the community of study. The Māori are indigenous people of Aotearoa (New Zealand). Prior to European settlement and colonial rule, the language of the Māori–Te Reo Māori–was a thriving language. The use of the Te Reo Māori started to decline as a consequence of colonial (language) policies and actions. For example, the Native Schools Act of 1867 established English as the sole language of instruction in schools [86], and resulted in the punishment of Māori children being punished for speaking Te Reo Māori in school [78].This, combined with events like epidemic of communicable diseases that reduce the population of the Māori resulted in an overall decline in the number of speakers of Te Reo Māori. With the complicated history of colonial rule, the Māori people have faced several obstacles in preserving and using their language. However, the community has actively fought for their language over the years, including in the implementation of policies [102], increasing access to language learning [86], and in raising awareness [78]. In 1972, the Māori Langauge Petition collected signatures from 30,000 people to urge the government of New Zealand to support the teaching of Te Reo Māori in schools [78]. In 1987, the Māori Language Act recognized Te Reo Māori as an official language of New Zealand, leading to efforts targeted towards the preservation and active use of the language [102]. Based on the existing Māori customs and protocols, Te Hiku Media more recently community-sourced annotated speech data for Te Reo Māori, and built several language technology tools for their community [43]

Listing the Capability Set for each stakeholder. Owing to the historical context provided above, the restrictions on the Te Reo Māori language mean that there is limited data and documentation about the language. Te Hiku Media sought to alleviate this problem and restore their language by collecting data to preserve Te Reo Māori and building language technologies for their community. Te Hiku Media relied on community sourcing their data through a competition that rewarded people for reading sentences in Te Reo Māori [22]. Te Hiku Media also had access to audio archives from their radio broadcasting service, which particularly had content relevant to the community, such as idiomatic phrases [22, 44]. As a community-rooted research initiative, Te Hiku Media had a unique position in the data collection landscape that allowed them to capture culturally relevant data for the Māori communities.

Listing the Conversion Factors for each stakeholder. Building language technologies requires infrastructure for hosting datasets and models, and computational resources to train and deploy models. Te Hiku Media demonstrated that with community-sourced datasets and the data they had already collected as a radio station, that they could train speech recognition models with error rates comparable to those available for English [22]. The Te Hiku team were able to win a grant that allowed them to further develop other language technologies [22]. Te Hiku Media also developed their own data license, which ensures that any partnership is conditioned on there being direct benefit to the Māori and that the data belongs to the Māori communities [22, 43]. The conversion factors that were necessary for the construction of the Te Hiku NLP project thus spanned technical and legal infrastructure, while the conversion factors necessary for the development of the system included social and technical measures for the Māori community to be able to donate data. These technical, legal, financial, and social infrastructures were necessary conversion factors due to the history of oppression, marginalization, externalization, and exclusion of the Māori community and language. Thus, Te Hiku Media laid out financial and legal protections for the Māori and their language, ensuring the Māori have the *real freedom* to build and utilize language technologies for Te Reo Māori.

Analysis using the Capabilities Approach. Te Hiku Media set out on a mission to preserve their language while ensuring community agency and direct community benefit are at the forefront of their endeavor. With decades of colonial policies that resulted in the decline in use of Te Reo Māori, Te Hiku Media paved a way for the Māori community to enjoy the benefits of language technologies while their data sovereignty and rights were protected. Furthermore, the

Te Hiku story has been used as a case study in prior work to demonstrate the reciprocity in the participatory work of Te Hiku in ensuring that the data, which was collected by the Māori, is used in projects where the Māori are the decision makers [14]. Thus, the Te Hiku story has also functioned as a source of inspiration for participatory research in language technologies. With the Capabilities Approach, we extend this discussion further to demonstrate how community-rooted research affords *real freedoms* to community members. Through their community-rooted efforts, Te Hiku Media ensured that the Māori have the legal and financial *conversion factors* needed to ensure the community's ultimate decision-making power in how–and by whom–their languages are used.

4.3 Case 3: Content Moderation in Diverse Socio-Cultural Contexts

In our last case study, we will offer a practical analysis that demonstrates how the CA allows us to disentangle the diverse intersectional identities of a given community when applied as an assessment tool. In particular, we will focus on failures of content moderation for low-resourced languages, which as prior work has demonstrated, are particularly pronounced for Majority World communities [98, 99]. We build this case study on the work of Nigatu and Raji [68], who investigated how users who search for Amharic content on YouTube are exposed to policy-violating content due to failures in content moderation.

Identifying the community of study. The first step in applying the Capabilities Approach is to identify the community we are studying and, in particular, what social structures are in place. The findings from Nigatu and Raji [68] demonstrate that Ethiopian women in particular are exposed to online harm due to (1) the failures of human and automated content moderation systems for low-resourced languages, and (2) malicious content creators who circumvent content moderation by, for instance, posing as medical doctors when they are posting policy-violating sexual content. Looking closely at the women who are affected by the policy violating content in the findings of Nigatu and Raji [68], we identify three personas: female migrant domestic worker from Ethiopia living in Middle Eastern countries, Ethiopian women living in Ethiopia, and Ethiopian women working white-collar jobs and studying in the United States and elsewhere in the diaspora. These women share in common their country of origin, the languages they speak, and how supported their languages are in recent advances in language technology research but diverge in terms of their geography and socio-economic standing.

Listing the Capability Set for each stakeholder. As stated above in Section 3.2, a capability set refers to the set of capabilities that are available to an individual. In Nigatu and Raji [68], the particular resource of focus is a social media platform, namely YouTube. Social media platforms serve as a source of entertainment, education, and even as a source of income for people across the globe [19, 82, 103, 112]. Findings from Nigatu and Raji [68] also suggest that individuals use the platform for religious and spiritual purposes and to access information about medical topics. Yet different stakeholders and user groups may still have different capabilities that impact how they use a platform. Here, we outline the capabilities of each of our three personas: First, the capability set is solely reliant on the freedom of each stakeholder to choose what they want to do. For example, a person may choose not to adopt or use the platform. Second, not all users may have equal access to the platform: some Ethiopian women living in the country may have less access to internet services, thus restricting their ability to utilize the platform. Third, given that the platform is exposing the users to harm, we may ask what other capabilities are available for each individual to access the services they seek on the platform (e.g. religious content, medical information) through alternate channels. For instance, the migrant domestic workers may not easily have access to medical resources and services [16, 41]. Further, prior work has found that migrant domestic workers who have been subject to abuse may be disinclined from seeking medical or professional

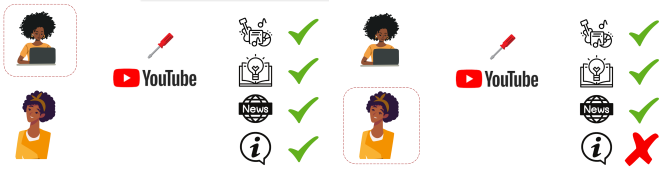

Fig. 1. While fixing the content moderation system may resolve all the issues for the Ethiopian software developer, it obfuscates the challenge with access to healthcare faced by the migrant domestic worker.

help, due to fears of retaliation by their abusers and to avoid legal complications related to their immigration status [26]. Instead, they may turn to community-oriented organizations and resources that migrant domestic workers have access to and that are sensitive to their situation, such as institutions started by migrant domestic workers themselves that are tailored towards providing migrant domestic workers with information and services [e.g., 1].

Listing the Conversion Factors for each stakeholder. Conversion factors depend on social (policies, norms, practices), personal (gender, physical ability), and environmental (buildings, communication, transport, climate, pollution, natural disasters) aspects of an individual's life [90]. Listing out the conversion factors allows us to reflect on what is and is not possible for each stakeholder, and what is preventing them from realizing their capabilities. In other words, it grounds our evaluation of what is *actually* possible for the community members. In our case study, conversion factors may include freely being able to access to the internet and digital devices, access to religious services, and access to healthcare facilities. Considering the three personas of our case study, each of them may have varying conversion factors. For instance, the Ethiopian women living in the country may have limited access to the internet, but might have more abundant access to Ethiopian religious services by virtue of being in the country.As a result, they might rely less on the platform as a source for religious content, or at the very least have the real freedom to access such content directly from religious institutions. On the other hand, Ethiopian women outside of the country may have limited options for accessing religious services, constraining their freedom to access such content without relying on the platform.

Analysis using the Capabilities Approach. In this case study, applying the Capabilities Approach sheds light on the diverse socio-economic standings and contexts of the three stakeholders under study. After evaluating the harms faced by Ethiopian women interacting with Amharic YouTube content, we might seek to propose several ways to address the moderation pipeline, e.g., by developing resources for local content moderation infrastructures that take into account the diverse challenges, co-developing support organizations, and developing policies the needs of each stakeholder [e.g. 98]. However, applying the Capabilities Approach and fully addressing the identified issues with content moderation infrastructures does not necessarily mean that the stakeholders will all be able to fully use the resource effectively. As we show in Figure 1, if we have a 'perfect' content moderation system for Amharic, the Ethiopian woman living in the US might be able to enjoy the benefits of the social media platform. However, the migrant domestic worker will still not have access to medical information, exposing her to harm via the lack of information. Thus, a 'perfect' content moderation system may obfuscate forms of harm that individuals in the community may face when using the technological artifact, conditioned on their socio-economic realities. With the Capabilities Approach, we were able to explicate the diverse social standings of the community of study and identify the blind spot of our evaluation and proposed solution.

5 Discussion

In this paper, we argue for adopting the Capabilities Approach framework for how we evaluate our processes for building language technologies. Particularly, we have looked at evaluating language technologies for the Majority World through the capabilities approach, where our current evaluation mechanisms do not account for the diverse cultures and languages and the long complicated history of colonization of the people of the Majority (Section 2). We first presented the theoretical background for the Capabilities Approach and demonstrated how we can adopt the framework to a language technology design and machine learning context (Section 3). We then used three case studies to demonstrate how the Capabilities Approach can be used to asses language technology design and evaluation, particularly highlighting its benefits (Section 4).

The core question of the Capabilities Approach is "What resources is a community capable of utilizing?" As the case study we presented in Section 4.1 illustrates, by prioritizing respect for community values and working with data that is available instead of attempting to match the STOA, Pine et al. [79] built efficient speech synthesis systems. In our case study in Section 4.2, we show how the Capabilities Approach allows us to effectively argue for a community-rooted approach to research in language technologies. With a Capabilities Approach lens, we interrogated how the impacts of colonial rule have restricted what capabilities were available to the Māori. In this case study, we find that the community-rooted approach to building language technology set the groundwork for financial and legal capability sets that allow the community to effectively resist exploitation and ensure direct benefit to the Māori. Our work connects to and extends prior work in participatory design [14], by bringing into conversation the real freedoms that a community-rooted research approach afforded the community.

The Capabilities Approach is particularly beneficial in highlighting how our interventions–such as language technologies–may further marginalize certain members of a community. In Section 4.3, we used content moderation in low-resourced languages as a case study to demonstrate how the socio-economic status of different users affects how effective a solution is towards a given population. We argued that, even if we theoretically fixed the content moderation problem, some users will still be exposed to (a different type) of digital harm. We further demonstrated how a 'perfect' content moderation system may in fact obstruct our view from identifying and intervening in harm directed at those with limited access to critical services such as healthcare.

Adopting a Capabilities Approach to evaluation would also mean questioning what language technologies we deem as useful for a community from the get-go. With the rise in Large Language Model (LLM) research, several works have geared towards evaluating [e.g. 73], adopting [e.g. 30, 46], and collecting data [e.g. 28] for low-resourced languages in relation to LLMs. With the Capabilities Approach in mind, we encourage language technology researchers to ask how effective such a tool would be for the communities who speak these languages. LLMs rely on large amounts of (textual) training data, which is not readily available for low-resourced languages. In collecting such data, the case study from Section 4.2 could help us map what mechanisms need to be in place to ensure the community members have the right conversion factors to protect themselves and their data from exploitation. Further, communities of the Majority have a rich culture of preserving knowledge through oral traditions [51]. With the Capabilities Approach, we would question whose stories we would exclude when employing a data collection scheme that solely relies on written records.

In Section 4 we have primarily provided examples of how the Capabilities Approach can be used as a method of evaluating and assessing initiatives and their impact. Yet, the Capabilities Approach also offers a framework for planning interventions by relying heavily on active involvement of the community and individuals for whom a technology is envisioned in the conceptualization and development process. The Capabilities Approach, as a planning tool is highly related to principles such as design justice [23], which emphasize the expertise of individuals and communities of their own lives and what will benefit them. When applying the Capabilities Approach in planning stages of a project, methods such as co-creation workshops [42] can afford individuals and communities with agency to intervene in researchers' and developers' assumptions about the resources that are available to the community, which existing capabilities they have, why they have (not) been converted into functioning(s), and capabilities that are desired by the community and the community's needs and prioritize with regard to technologies.

Limitations. As we have demonstrated in this paper, adopting the Capabilities Approach into how we design and evaluate language technologies would illuminate gaps due to social structures. However, it is not without limitations. First, we have to be careful in the assumptions we make as researchers when adopting the approach. One way to mitigate importing our researcher bias into how we define capabilities for a community is to use the procedural approach [81, 94]. The procedural approach emphasizes community ownership over determining what is important and avoids researchers putting subjective constraints. This is particularly important in Majority World contexts, where research often does not value the voices of community members and usually provides a deficit-oriented framing. Second, the Capabilities Approach, when adopted as we are proposing it in this paper, is a diagnostic tool for assessing our design and evalaution processes. While it will help us identify blind spots in our evaluation, it will not necessarily lead to solutions by itself. For instance, understanding that a 'perfect' content moderation system will not protect all users from harm in our case study (Section 4.3) requires us to go a step further to resolve the challenge faced by those specific users. Often, the solutions may require cross-disciplinary collaborations, as the issues may stem from social, economic, or political sources.

6 Conclusion

In this paper, we have proposed adopting the Capabilities Approach to how we evaluate language technologies. The Capabilities Approach encourages us to shift from merely questions about theoretical technical performance, and instead asks "what resources does a community have the capability to utilize?" in a challenge to our implicit assumptions about resource allocation. We have demonstrated how adopting a Capabilities Approach lens to our evaluation can help us foreground the diverse identities of the Majority World and how those identities impact the utility of technological solutions. Through three case studies, we have illustrated how one can adopt the Capabilities Approach to language technology evaluation in practice and have highlighted the benefits of doing so. In our discussion, we also consider the Capabilities Approach as a tool for planning the development of language technologies and resources. While the Capabilities Approach has been applied in a number of different fields and areas, future work is required to identify how to apply it within planning, which risks and concerns arise, and how to navigate the political and social landscape in the development. Our work thus serves as an argument for the importance and benefits of meaningful inclusion of languages and communities of the Majority in research–one that centers their needs, capabilities, and agency, and we hope it can serve as an inspiration future work in carefully and thoughtfully building language technologies for understudied languages.

Generative AI statement

In this paper, we have not used any form of generative artificial intelligence or other machine learning tools, outside of machine learning systems embedded in traditional search interfaces.

References

[1] 2025. Egna Legna Besidet. https://egnalegna.org/about-us-depricated

[2] 2025. Mohawk Language (MOH) – L1 & L2 Speakers, Status, Map, Endangered Level & Official Use | Ethnologue Free. https://www.ethnologue.com/language/moh

[3] 2025. Onkwawenna Kentyohkwa. https://onkwawenna.info

[4] Rediet Abebe, Kehinde D. Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou Lionel Remy, and Swathi Sadagopan. 2021. Narratives and Counternarratives on Data Sharing in Africa. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021). https://api.semanticscholar.org/CorpusID:232040609

[5] Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying Sentiments in Algerian Code-switched User-generated Comments. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2698–2705. https://aclanthology.org/2020.lrec-1.328/

[6] Jesujoba O. Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT Corpus for African Languages. arXiv:2501.06374 [cs.CL] https://arxiv.org/abs/2501.06374

[7] Shahidul Alam. 2008. Majority World: Challenging the West's Rhetoric of Democracy. https://www.researchgate.net/publication/285046329_Majority_World_Challenging_the_West%27s_Rhetoric_of_Democracy

[8] Hizkiel Mitiku Alemayehu, Hamada M Zahera, and Axel-Cyrille Ngonga Ngomo. 2024. Error Analysis of Multilingual Language Models in Machine Translation: A Case Study of English-Amharic Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 19758–19768. https://doi.org/10.18653/v1/2024.emnlp-main.1102

[9] Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni, and Jackie Chi Kit Cheung. 2023. Varta: A Large-Scale Headline-Generation Dataset for Indic Languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3468–3492. https://doi.org/10.18653/v1/2023.findings-acl.215

[10] Mercy Nyamewaa Asiedu, Awa Dieng, Iskandar Haykel, Negar Rostamzadeh, Stephen Pfohl, Chirag Nagpal, Maria Nagawa, Abigail Oppong, Sanmi Koyejo, and Katherine Heller. 2024. The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, Article 10, 24 pages. https://doi.org/10.1145/3689904.3694708

[11] David Barasa. 2023. Language ideologies, policies and practices within the multilingual Kenyan context. *JLLCS* 2, 1 (Sept. 2023), 55–62. https://doi.org/10.58721/jllcs.v2i1.336

[12] Steven Bird. 2024. Must NLP be Extractive?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 14915–14929. https://doi.org/10.18653/v1/2024.acl-long.797

[13] Abeba Birhane. 2020. Algorithmic Colonization of Africa. *SCRIPTed: A Journal of Law, Technology and Society 17 SCRIPT* 17, 2 (2020). https://doi.org/10.2966/scrip.170220.389

[14] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3551624.3555290

[15] Abeba Birhane and Zeerak Talat. 2023. It's incomprehensible: on machine learning and decoloniality. In *Handbook of Critical Studies of Artificial Intelligence*, Simon Lindgren (Ed.). Edward Elgar Publishing, 128–140. https://doi.org/10.4337/9781803928562.00016

[16] Lisa Blaydes. 2023. Assessing the Labor Conditions of Migrant Domestic Workers in the Arab Gulf States. *ILR Review* 76, 4 (Jan. 2023), 724–747. https://doi.org/10.1177/00197939221147497

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

[18] Adam Bouyamourn. 2023. Why LLMs Hallucinate, and How to Get (Evidential) Closure: Perceptual, Intensional, and Extensional Learning for Faithful Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3181–3193. https://doi.org/10.18653/v1/2023.emnlp-main.192

[19] Hung Phu Bui, Mark Bedoya Ulla, Veronico N. Tarrayo, and Chien Thang Pham. 2023. Editorial: The roles of social media in education: affective, behavioral, and cognitive dimensions. *Front. Psychol.* 14 (Sept. 2023), 1287728. https://doi.org/10.3389/fpsyg.2023.1287728

[20] Garance Burke and Hilke Schellmann. 2024. Researchers say AI transcription tool used in hospitals invents things no one ever said. *AP News* (Oct. 2024). https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14

[21] Latika Chaudhary. 2007. An Economic History of Education in Colonial India. https://scispace.com/papers/an-economic-history-of-education-in-colonial-india-33phxv1ipt

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

[23] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* MIT Press. Google-Books-ID: m4LPDwAAQBAJ.

[24] Karina Czyzewski. 2011. Colonialism as a Broader Social Determinant of Health. *iipj* 2, 1 (May 2011). https://doi.org/10.18584/iipj.2011.2.1.5

[25] Amol S. Dhane, Sachin Sarode, Gargi Sarode, and Shruti Singh. 2024. A reality check on chatbot-generated references in global health research. *Oral Oncology Reports* 10 (June 2024), 100246. https://doi.org/10.1016/j.oor.2024.100246 [Online; accessed 2. May 2025].

[26] Jasmin Lilian Diab, Banchi Yimer, Tsigereda Birhanu, Ariane Kitoko, Amira Gidey, and Francisca Ankrah. 2023. The gender dimensions of sexual violence against migrant domestic workers in post-2019 Lebanon. *Frontiers in Sociology* 7 (Jan. 2023), 1091957. https://doi.org/10.3389/fsoc.2022.1091957

[27] Harshita Diddee, Kalika Bali, Monojit Choudhury, and Namrata Mukhija. 2022. The Six Conundrums of Building and Deploying Language Technologies for Social Good. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies* (Seattle, WA, USA) *(COMPASS '22)*. Association for Computing Machinery, New York, NY, USA, 12–19. https://doi.org/10.1145/3530190.3534792

[28] Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 12402–12426. https://doi.org/10.18653/v1/2023.acl-long.693

[29] andy Dong. 2008. The Policy of Design: A Capabilities Approach on JSTOR. , 76–87 pages. https://www.jstor.org/stable/25224195

[30] Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In *Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, Angela Fan, Iryna Gurevych, Yufang Hou, Zornitsa Kozareva, Sasha Luccioni, Nafise Sadat Moosavi, Sujith Ravi, Gyuwan Kim, Roy Schwartz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 52–64. https://doi.org/10.18653/v1/2022.sustainlp-1.11

[31] Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4820–4831. https://doi.org/10.18653/v1/2020.emnlp-main.391

[32] Ethnologue. 2025. *What continents have the most indigenous languages?* https://www.ethnologue.com/insights/continents-most-indigenous-languages

[33] Meng Fang and Trevor Cohn. 2017. Model Transfer for Tagging Low-resource Languages using a Bilingual Dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 587–593. https://doi.org/10.18653/v1/P17-2093

[34] Isaac Feldman and Rolando Coto-Solano. 2020. Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 3965–3976. https://doi.org/10.18653/v1/2020.coling-main.351

[35] Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. 2023. Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11857–11870. https://doi.org/10.18653/v1/2023.acl-long.661

[36] Jeremy Green. 2018. Kanyen'kéha: Mohawk Language. *The Canadian Encyclopedia* (May 2018). https://thecanadianencyclopedia.ca/en/article/kanyenkeha-mohawk-language#IntroductionTheKanyenkehka

[37] Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1926–1939. https://doi.org/10.1145/3630106.3659017

[38] Saiful Haq, Ashutosh Sharma, Omar Khattab, Niyati Chhaya, and Pushpak Bhattacharyya. 2024. IndicIRSuite: Multilingual Dataset and Neural Information Models for Indian Languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 501–509. https://doi.org/10.18653/v1/2024.acl-short.46

[39] William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A Material Lens on Coloniality in NLP. arXiv:2311.08391 [cs.CL] https://arxiv.org/abs/2311.08391

[40] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behav. Brain Sci.* 33, 2-3 (June 2010), 61–83. https://doi.org/10.1017/S0140525X0999152X arXiv:20550733

[41] Amnesty International. 2021. Lebanon: 'Their house is my prison': Exploitation of migrant domestic workers in Lebanon - Amnesty International. https://www.amnesty.org/en/documents/mde18/0022/2019/en

[42] Peter Jones. 2018. Contexts of Co-creation: Designing with System Stakeholders. In *Systemic Design*, Peter Jones and Kyoichi Kijima (Eds.). Vol. 8. Springer Japan, Tokyo, 3–52. https://doi.org/10.1007/978-4-431-55639-8_1 Series Title: Translational Systems Sciences.

[43] Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2023. Kia tangata whenua: Artificial intelligence that grows from the land and people. *Ethical Space: International Journal of Communication Ethics* 2023, 2/3 (Aug. 2023). https://doi.org/10.21428/0af3f4c0.9092b177

[44] Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2023. Ngā taonga tuku iho: Intergenerational transmission using archives. *Ethical Space: International Journal of Communication Ethics* 2023, 2/3 (Aug. 2023). https://doi.org/10.21428/0af3f4c0.43228e07

[45] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[46] Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?Small Data? Pretraining Approaches to Sequence-to-Sequence Models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta (Eds.). Association for Computational Linguistics, Hybrid, 126–135. https://doi.org/10.18653/v1/2022.deeplo-1.14

[47] Rachael Ka'ai-Mahuta. 2011. The impact of colonisation on te reo Māori: A critical review of the State education system. *Te Kaharoa* 4 (12 2011). https://doi.org/10.24135/tekaharoa.v4i1.117

[48] Themrise Khan, Seye Abimbola, Catherine Kyobutungi, and Madhukar Pai. 2022. How we classify countries and people—and why it matters. *BMJ Global Health* 7, 6 (June 2022), e009704. https://doi.org/10.1136/bmjgh-2022-009704

[49] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1672–1681. https://doi.org/10.1145/3630106.3658996

[50] Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2022. Out of Thin Air: Is Zero-Shot Cross-Lingual Keyword Detection Better Than Unsupervised?. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 400–409. https://aclanthology.org/2022.lrec-1.42/

[51] Lindah Kotut and D. Scott McCrickard. 2022. Winds of Change: Seeking, Preserving, and Retelling Indigenous Knowledge Through Self-Organized Online Communities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 257, 15 pages. https://doi.org/10.1145/3491102.3502094

[52] Tahu Kukutai, Shemana Cassim, Vanessa Clark, Nicholas Jones, Jason Mika, Rhianna Morar, Marama Muru-Lanning, Robert Pouwhare, Vanessa Teague, Lynell Tuffery Huria, David Watts, and Rogena Sterling. 2023. Māori data sovereignty and privacy. *Te Ngira Institute for Population Research* (2023). https://researchcommons.waikato.ac.nz/items/8b16bd0e-bf41-4630-a721-3f343912b69d

[53] Clive Lawson. 2010. Technology and the Extension of Human Capabilities. *Journal for the Theory of Social Behaviour* 40, 2 (June 2010), 207–223. https://doi.org/10.1111/j.1468-5914.2009.00428.x

[54] Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8608–8621. https://doi.org/10.18653/v1/2022.emnlp-main.590

[55] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

[56] Natasha Ita MacDonald. 2023. Why Inuit culture and language matter: decolonizing English second language learning. *AlterNative: An International Journal of Indigenous Peoples* (dec 2023). https://doi.org/10.1177/11771801231197841

[57] Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 521–534. https://doi.org/10.1145/3531146.3533117

[58] Aphile-amanzima Mazibuko. 2023. Africa's Digital Gender Divide – ACCORD. https://www.accord.org.za/analysis/africas-digital-gender-divide

[59] Alamin Mazrui and Ali Mazrui. 2025. Dominant Languages in a Plural Society: English and Kiswahili in Post-Colonial East Africa on JSTOR. https://www-jstor-org.libproxy.berkeley.edu/stable/1601194?seq=1

[60] J. A. Meaney, Beatrice Alex, and William Lamb. 2024. Testing and Adapting the Representational Abilities of Large Language Models on Folktales in Low-Resource Languages. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni (Eds.). Association for Computational Linguistics, Miami, USA, 319–324. https://doi.org/10.18653/v1/2024.nlp4dh-1.31

[61] Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2016–2025. https://doi.org/10.1145/3531146.3533244

[62] Bettina Migge and Isabelle Léglise. 2007. *Language and colonialism. Applied linguistics in the context of creole communities.* 297–338. https://doi.org/10.1515/9783110198539.2.299

[63] Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L

Lopez-Davila, Maraim Masoud, Nikita Nangia, Anaelia Ovalle, Giada Pistilli, Dragomir Radev, Beatrice Savoldi, Vipul Raheja, Jeremy Qin, Esther Ploeger, Arjun Subramonian, Kaustubh Dhole, Kaiser Sun, Amirbek Djanibekov, Jonibek Mansurov, Kayo Yin, Emilio Villa Cueva, Sagnik Mukherjee, Jerry Huang, Xudong Shen, Jay Gala, Hamdan Al-Ali, Tair Djanibekov, Nurdaulet Mukhituly, Shangrui Nie, Shanya Sharma, Karolina Stanczak, Eliza Szczechla, Tiago Timponi Torrent, Deepak Tunuguntla, Marcelo Viridiano, Oskar Van Der Wal, Adina Yakefu, Aurélie Névéol, Mike Zhang, Sydney Zink, and Zeerak Talat. 2025. SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 11995–12041. https://aclanthology.org/2025.naacl-long.600/

[64] Evelyne Musambi and Cara Anna. 2023. Facebook content moderators in Kenya call the work 'torture.' Their lawsuit may ripple worldwide. *AP News* (June 2023). https://apnews.com/article/kenya-facebook-content-moderation-lawsuit-8215445b191fce9df4ebe35183d8b322

[65] Stanslaus Mwongela, Jay Patel, Sathy Rajasekharan, Laura Wotton, Mohammed Ahmed, Gilles Hacheme, Bernard Shibwabo, and Julius Butime. 2023. Data-Efficient Learning For Healthcare Queries In Low-Resource and Code Mixed Language Settings. *Practical ML for Developing Countries Workshop at the International Conference on Learning Representations (ICLR)* (2023).

[66] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 16366–16393. https://doi.org/10.18653/v1/2024.acl-long.862

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

[68] Hellina Hailu Nigatu and Deborah Inioluwa Raji. 2024. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. *FAccT* (June 2024). https://drive.google.com/file/d/1aSohyoaaeQxoZ89gT3sa8fabRyK9pZex/view

[69] Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's Paradox of 'Low-Resource' Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17753–17774. https://doi.org/10.18653/v1/2024.emnlp-main.983

[70] Rick Noack. 2015. The world's languages, in seven maps and charts | The Independent. *Independent* (Dec. 2015). https://www.independent.co.uk/news/world/the-world-s-languages-in-seven-maps-and-charts-a6791871.html

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

[72] Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are Large Language Models on African Languages? *arXiv* (Nov. 2023). https://doi.org/10.48550/arXiv.2311.07978 arXiv:2311.07978

[73] Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? arXiv:2311.07978 [cs.CL] https://arxiv.org/abs/2311.07978

[74] Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics* 11 (2023), 1669–1685. https://doi.org/10.1162/tacl_a_00627

[75] Ilse Oosterlaken. 2009. Design for Development: A Capability Approach. *Design Issues* 25, 4 (2009), 91–102. http://www.jstor.org/stable/20627832

[76] Ilse Oosterlaken. 2011. Inserting Technology in the Relational Ontology of Sen's Capability Approach. *Journal of Human Development and Capabilities* (Aug. 2011). https://www.tandfonline.com/doi/full/10.1080/19452829.2011.576661#d1e330

[77] Ilse Oosterlaken. 2015. Human Capabilities in Design for Values. In *Handbook of Ethics, Values, and Technological Design*. Springer, Dordrecht, The Netherlands, 221–250. https://doi.org/10.1007/978-94-007-6970-0_7

[78] New Zeland Parliament. [n. d.]. *Te Petihana Reo Māori — The Māori Language Petition - New Zealand Parliament*. https://www.parliament.nz/en/visit-and-learn/history-and-buildings/te-rima-tekau-tau-o-te-petihana-reo-maori-the-50th-anniversary-of-the-maori-language-petition/te-petihana-reo-maori-the-maori-language-petition/

[79] Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 7346–7359.

https://doi.org/10.18653/v1/2022.acl-long.507

[80] Manas Ranjan Pradhan and Prasenjit De. 2025. Women’s healthcare access: assessing the household, logistic and facility-level barriers in India. *BMC Health Serv. Res.* 25 (Feb. 2025), 323. https://doi.org/10.1186/s12913-025-12463-9

[81] Amartya Sen Lamont University Professor. 2005. Human Rights and Capabilities. *Journal of Human Development* (July 2005). https://www-tandfonline-com.libproxy.berkeley.edu/doi/full/10.1080/14649880500120491#d1e177

[82] Amanda Raffoul, Zachary J. Ward, Monique Santoso, Jill R. Kavanaugh, and S. Bryn Austin. 2023. Social media platforms generate billions of dollars in revenue from U.S. youth: Findings from a simulated revenue model. *PLoS One* 18, 12 (Dec. 2023), e0295337. https://doi.org/10.1371/journal.pone.0295337

[83] Massimo Ragnedda. 2019. *Conceptualising the digital divide.* Amsterdam University Press, 27–44. http://www.jstor.org/stable/j.ctvh4zj72.6

[84] Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessico Ojo, Atnafu Lambebo Tonja, Maushami Chetty, Onyothi Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2025. The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages. arXiv:2502.15916 [cs.CL] https://arxiv.org/abs/2502.15916

[85] Torsten Masson Rebecca Gutwald, Ortrud Leßmann and Felix Rauschmayer. 2014. A Capability Approach to Intergenerational Justice? Examining the Potential of Amartya Sen’s Ethics with Regard to Intergenerational Issues. *Journal of Human Development and Capabilities* 15, 4 (2014), 355–368. https://doi.org/10.1080/19452829.2014.899563 arXiv:https://doi.org/10.1080/19452829.2014.899563

[86] Elaine Reese, Peter Keegan, Stuart Mcnaughton, Te Kani Kingi, Polly Atatoa Carr, Johanna Schmidt, Jatender Mohal, Cameron Grant, and Susan Morton. 2018. Te Reo Māori: indigenous language acquisition in the context of New Zealand English∗. *J. Child Lang.* 45, 2 (March 2018), 340–367. https://doi.org/10.1017/S0305000917000241

[87] Dr. Salvador Santino F. Regilme. 2025. Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South. *sais* 44, 2 (Feb. 2025), 75–92. https://doi.org/10.1353/sais.2024.a950958

[88] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI ’22)*. Association for Computing Machinery, New York, NY, USA, Article 299, 17 pages. https://doi.org/10.1145/3491102.3517639

[89] Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Nina Markl, Léa-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2023. Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI ’23)*. Association for Computing Machinery, New York, NY, USA, Article 406, 17 pages. https://doi.org/10.1145/3544548.3581385

[90] Ingrid Robeyns and Morten Fibieger Byskov. 2011. The Capability Approach. https://plato.stanford.edu/entries/capability-approach

[91] Renato Rocha Souza, Amelie Dorn, Barbara Piringer, and Eveline Wandl-Vogt. 2020. Identification of Indigenous Knowledge Concepts through Semantic Networks, Spelling Tools and Word Embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 943–947. https://aclanthology.org/2020.lrec-1.118/

[92] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili’c, Daniel Hesslow, Roman Castagn’e, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz’alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Moham mad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L’opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank

Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Dan Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenˇek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, José D. Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, Patrick Haller, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *ArXiv* abs/2211.05100 (2022). https://api.semanticscholar.org/CorpusID:253420279

[93] David Schlosberg. 2012. Climate Justice and Capabilities: A Framework for Adaptation Policy. *Ethics &; International Affairs* 26, 4 (2012), 445–461. https://doi.org/10.1017/S0892679412000615

[94] Amartya Sen. 1974. Informational bases of alternative welfare approaches: Aggregation and income distribution. *Journal of Public Economics* 3, 4 (nov 1974), 387–403. https://doi.org/10.1016/0047-2727(74)90006-1

[95] Amartya Sen. 1983. Poor, Relatively Speaking on JSTOR. , 153–169 pages. https://www.jstor.org/stable/2662642

[96] Amartya Sen. 1999. Development as Freedom. *Oxford University Press.* (1999).

[97] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 160–171. https://doi.org/10.1145/3593013.3593985

[98] Farhana Shahid, Mona Elswah, and Aditya Vashistha. 2025. Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages. arXiv:2501.13836 [cs.CL] https://arxiv.org/abs/2501.13836

[99] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. https://doi.org/10.1145/3544548.3581538

[100] Matthew L. Smith and Carolina Seward. 2009. The Relational Ontology of Amartya Sen's Capability Approach: Incorporating Social and Individual Causes. *Journal of Human Development and Capabilities* (July 2009). https://www.tandfonline.com/doi/10.1080/19452820902940927#abstract

[101] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (Eds.). Association for Computational Linguistics, virtual+Dublin, 26–41. https://doi.org/10.18653/v1/2022.bigscience-1.3

[102] Manatū Taonga. 2024. History of the Māori language. https://nzhistory.govt.nz/culture/maori-language-week/history-of-the-maori-language

[103] Hedviga Tkacová, Roman Králik, Miroslav Tvrdoň, Zita Jenisová, and José García Martin. 2022. Credibility and Involvement of Social Media in Education—Recommendations for Mitigating the Negative Effects of the Pandemic among High School Students. *Int. J. Environ. Res. Public Health* 19, 5 (Feb. 2022), 2767. https://doi.org/10.3390/ijerph19052767

[104] Elaine Unterhalter. 2005. Global inequality, capabilities, social justice: The millennium development goal for gender equality in education. *International Journal of Educational Development* 25, 2 (March 2005), 111–122. https://doi.org/10.1016/j.ijedudev.2004.11.015

[105] Robin Vandecasteele, Lenzo Robijn, Sara Willems, Stéphanie De Maesschalck, and Peter A. J. Stevens. 2024. Barriers and facilitators to culturally sensitive care in general practice: a reflexive thematic analysis. *BMC Primary Care* 25 (Oct. 2024), 381. https://doi.org/10.1186/s12875-024-02630-y

[106] Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2203–2213. https://doi.org/10.18653/v1/2021.eacl-main.188

[107] Sridhar Venkatapuram. 2011. *Health Justice: An Argument From the Capabilities Approach*. Polity Press.

[108] Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Rebecca Knowles, Akiko Eriguchi, and Shivali Goel (Eds.). Association for Machine Translation in the Americas, Chicago, USA, 236–249. https://aclanthology.org/2024.amta-research.20/

[109] Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. 2024. Mitigating the Language Mismatch and Repetition Issues in LLM-based Machine Translation via Model Editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15681–15700. https://doi.org/10.18653/v1/2024.emnlp-main.879

[110] Mirriam Wangui. 2024. Impact of Colonial Policies on Indigenous Education Systems in Africa. *EJHR* 3, 2 (Aug. 2024), 32–45. https://doi.org/10.47672/ejhr.2334

[111] Daricia Wilkinson and Bart Knijnenburg. 2022. Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 102, 25 pages. https://doi.org/10.1145/3491102.3517643

[112] Dominique S. Wirz and Florin Zai. 2025. Infotainment on Social Media: How News Companies Combine Information and Entertainment in News Stories on Instagram and TikTok. *Digital Journalism* (Feb. 2025). https://doi.org/10.1080/21670811.2025.2464062

[113] Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking Machine Translation with Cultural Awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13078–13096. https://doi.org/10.18653/v1/2024.findings-emnlp.765

[114] Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-Resource Languages Jailbreak GPT-4. *arXiv* (Oct. 2023). https://doi.org/10.48550/arXiv.2310.02446 arXiv:2310.02446