IT'S A GLOBAL VILLAGE (IF YOU SPEAK THE RIGHT LANGUAGE): ON LANGUAGE MODELS, DIGITAL SIDELINING, AND PARTICIPATION

NOA MOR*

ABSTRACT

The digital linguistic ecosystem is rife with disparities. While a select group among the world's seven thousand languages enjoys the benefits of digitalization, speakers of Digitally Marginalized Languages (DMLs) have restricted or no access to these resources. Focusing on the AI fields of Natural Language Processing (NLP) and Large Language Models (LLMs), this Article explores the nature of these linguistic gaps and how they compound and exacerbate long-standing, offline linguistic hierarchies.

This Article analyzes the techno-social predicaments that underpin these inequalities, addressing two key areas: (1) training data and training processes, and (2) design and evaluation choices and constraints. Drawing on Nancy Fraser's "Parity of Participation" framework, this Article examines how these predicaments, along with cultural and regulatory criteria, prevent equal participation for DML speakers across three dimensions: distribution, recognition, and representation.

This Article then explores the international human rights law framework that applies to governments and private AI companies, and outlines their duties and responsibilities in facilitating DML speakers' participation. By tying together technological, fairness, and legal perspectives, this Article provides a comprehensive and novel look into global linguistic discrepancies in the digital age and how they can be tackled.

---

* Postdoctoral Fellow, Faculty of Law, The Hebrew University of Jerusalem. The author thanks Yohannes Eneyew Ayalew, Amir Cahane, Dafna Dror-Shpoliansky, Tamar Megiddo, Gadi Perl, Tomer Shadmy, Yuval Shany, Maria Varaki, Mia Or Winraob, and the participants of the EASST – 4S 2024 for their valuable comments. The author also thanks the WILJ editorial team for their exceptional contribution to this article. The research was conducted with the support of ERC Grant No. 101054745: Three Generations of Digital Human Rights.

https://doi.org/10.59015/wilj.SCTW4297

https://doi.org/10.59015/wilj.SCTW4297

INTRODUCTION

The digital world provides unparalleled opportunities and mechanisms for communication and information seeking, analyzing, and sharing.[1] Digital tools are now widely enmeshed in modern life, from social media, video sharing, search engines, digital wallets, e-commerce, and e-governance platforms, to fitness apps, productivity and education tools, workplace management software, and navigation products. These tools serve as a portal for many everyday activities and are fundamental to individuals' and communities' ability to exercise a wide range of human rights and access societal and political benefits.[2]

Despite these advantages, the ability to participate in digital avenues significantly varies among speakers of different languages. Of the world's approximately seven thousand languages,[3] only speakers of a select number of languages have meaningful access to dominant, everyday digital avenues and the advantages they afford. Conversely, speakers of the vast majority of the languages in the world only have limited access to these spaces or are excluded from them altogether.[4]

Discrepancies among languages, however, emerged long before digitalization. Such gaps were largely shaped by long-standing and powerful processes of colonialism, nationalism, and globalization.[5] Digitalization generated an intricate set of challenges that compound existing offline linguistic asymmetries and, at times, further deepen them. Indeed, more than a third of the world's population is unconnected to the

---

[1] *See generally* Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011 (2018); Andreas M. Kaplan & Michael Haenlein, *Users of the World, Unite! The Challenges and Opportunities of Social Media*, 53 BUS. HORIZONS 59 (2010).

[2] *See* Stephen Tully, *A Human Right to Access the Internet? Problems and Prospects*, 14 HUM. RIGHTS LAW REV. 175, 176–177 (2014). In the context of social networks*, see generally* Packingham v. North Carolina, 582 U.S. 98 (2017).

[3] *How many languages are there in the world?*, ETHNOLOGUE, https://www.ethnologue.com/insights/how-many-languages/ [https://perma.cc/5ELK-RUYZ] (last visited Nov. 7, 2023).

[4] *See infra* Part I.A for the discussion on this later in this Article. Some describe language as "a set of rules or set of symbols where symbols are combined and used for conveying information or broadcasting the information," Diksha Khurana et al., *Natural Language Processing: State of the Art, Current Trends and Challenges*, *in* 82 MULTIMEDIA TOOLS & APPLICATIONS 3713, 3714 (2022). Others embed the identity and cultural importance of the languages in its definition. *See, e.g.,* RAYMOND WILLIAMS, MARXISM & LITERATURE 21 (1978) ("A definition of language is always, implicitly or explicitly, a definition of human beings in the world.").

[5] *See* Thomas Hylland Eriksen, *Linguistic Hegemony and Minority Resistance*, 29 J. PEACE RECH. 313, 314-18 (1992); see *infra* Part I.B for the discussion on this later in this Article.

https://doi.org/10.59015/wilj.SCTW4297

internet.[6] Even when a connection exists, speakers on the wrong side of the linguistic digital map face other constraints like equipment-related difficulties and the limited number of supported languages, including in popular apps and services such as WhatsApp, Facebook, and YouTube.[7]

Additional manifestations of the digital linguistic gaps lie in the languages in which online content appears. English leads the list of online content and is present in roughly 50 percent of the global web content.[8] The following nine leading languages lag significantly behind English, but together with English, these ten languages account for around 85 percent of all the online content, despite representing but a small fraction of the world's languages.[9]

Indeed, popular services, such as Google Maps and Wikipedia, offer much more information in English and other digitally dominant languages, compared to less dominant languages.[10] Content gaps between these groups are not confined to issues of scope or density, though. When speakers of marginalized languages go online, they may find that existing informational resources fail to embody their collective histories, values, needs, and lived experiences. This partly stems from the fact that it is the speakers of digitally dominant languages that often tell the stories of communities of more vulnerable languages through the creation of online content.[11] Moreover, in cases where speakers of digitally marginalized languages create online content, certain problems emerge from within. The authors of such content were sometimes members of the strong subgroups in that given linguistic sector, resulting in the silencing of vulnerable

---

[6] *Summary Report*, STATE OF THE INTERNET'S LANGUAGES REPORT (2022), https://internetlanguages.org/en/summary/ [https://perma.cc/5W6W-HAYJ].

[7] *See Infra* Part I.A for the discussion on this later in this Article.

[8] Ani Petrosyan, *Languages most frequently used for web content as of January 2024, by share of websites*, STATISTA (Oct. 21, 2024), https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/ [https://perma.cc/88JA-EVBP].

[9] *Id.*; For different figures (in which the overall picture of asymmetry is maintained), *see* Usage Statistics of Content Languages for Websites, W3TECHS, https://w3techs.com/technologies/overview/content_language [https://perma.cc/D48G-FJUW] (last accessed July 3, 2024); *Summary Report*, *supra* note 6.

[10] *Summary Report*, *supra* note 6; The language geography of Google Maps, State of the Internet's Languages Report, https://internetlanguages.org/en/numbers/google-maps-language-geography/ [https://perma.cc/6K27-MYRR] (last accessed Nov 9, 2023); see also *infra* Part I.A for the discussion on this later in this Article.

[11] *Id.*

https://doi.org/10.59015/wilj.SCTW4297

voices, including those of the LGBTQ+ community, women, and persons with disabilities, within that sector.[12]

Such linguistic concerns also arise with emerging and popular AI technologies, particularly Natural Language Processing (NLP) and Large Language Models (LLMs).[13] The latter include transformative products such as OpenAI's Generative Pretrained Transformer (GPT), Google's Gemini, and Anthropic's Claude.[14]

These technologies introduce breakthrough opportunities for humanity in different contexts and fields.[15] They perform a wide range of complex linguistic tasks, such as question answering, summarization, information analysis and extraction, and writing assistance. Moreover, LLMs are being integrated into applications that span multiple facets of life—including health,[16] education,[17] law,[18] finance,[19] and disaster response[20]—thereby carrying an even broader societal impact. Among the benefits those technologies provide, significant possibilities are introduced

---

[12] *Summary Report*, *supra* note 6; Josia P. Darmawan, *Flickering Hope: Challenges in Creating Online LGBTQIA+ Content in Bahasa Indonesia*, STATE OF THE INTERNET'S LANGUAGES REPORT (2022), https://internetlanguages.org/en/stories/flickering-hope/ [https://perma.cc/CWA4-69FR].

[13] *See generally* Surangika Ranathunga & Nisansa de Silva, *Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World*, *in* 1 PROCEEDINGS OF THE 2ND CONFERENCE OF THE ASIA-PACIFIC CHAPTER OF THE ASSOCIATION FOR COMPUTATION LINGUISTICS AND THE 12TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING 823 (Yulan He & Heng Ji et al., eds. 2022).

[14] *See* Enkelejda Kasneci et al., *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*, 103 LEARNING AND INDIVIDUAL DIFFERENCES art. 102274, 2 (2023).

[15] See generally with relation to such opportunities: Viet Dac Lai et al., *ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning*, ARXIV, (Apr 12, 2023), http://arxiv.org/abs/2304.05613 [https://perma.cc/3XRY-FPQ4] (see PDF). For a discussion on the benefits provided to a small group of dominant languages in the context of LLMs, see *infra* Part II.B.2.b below. It should be noted that alongside the benefits afforded by LLMs, they also pose significant challenges, *inter alia,* regarding privacy, copyright, bias, and "hallucinations." *See, e.g.*, *id.* and Hadas Kotek et al., *Gender Bias and Stereotypes in Large Language Models in* CI '23: COLLECTIVE INTELLIGENCE CONFERENCE 12 (2023).

[16] *See generally* Mahyar Abbasian et al., *Conversational Health Agents: A Personalized LLM-Powered Agent Framework*, ARXIV, (Sept. 26, 2024), https://doi.org/10.48550/arXiv.2310.02374 [https://perma.cc/9GAB-HSPE] (see PDF).

[17] *See generally* Nursultan Askarbekuly & Nenad Aničić, *LLM Examiner: Automating Assessment in Informal Self-Directed E-Learning Using ChatGPT*, 66 KNOWLEDGE & INFO. SYS. 6133 (2024).

[18] *See generally* Jiaxi Cui et al., *Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model*, ARXIV (May 30, 2024), https://doi.org/10.48550/arXiv.2306.16092 [https://perma.cc/3HHA-36F4] (see PDF).

[19] *See generally* Shijie Wu et al., *BloombergGPT: A Large Language Model for Finance*, ARXIV (Dec. 21, 2023), https://doi.org/10.48550/arXiv.2303.17564 [https://perma.cc/2THV-NSEV] (see PDF).

[20] Ranathunga & de Silva, *supra* note 13, at 823 (in the context of NLP in general).

https://doi.org/10.59015/wilj.SCTW4297

to speakers of some marginalized languages, including machine translation, content generation, and access to digital services.[21] Indeed, valuable efforts are made by various stakeholders to enhance the multilingual capabilities of LLMs, and to support the development of language-specific models that cater to a broader range of languages.[22]

Notwithstanding these advancements, current LLMs leave behind the bulk of the world's languages or provide them with limited and low-performance solutions compared to those available to the speakers of English and several other dominant languages.[23]

Currently, LLMs provide far-reaching, transformative benefits for (already) dominant languages, while only offering limited, albeit valuable, advantages to a small segment of the remaining languages. This indicates that LLMs not only embody existing linguistic gaps but may further exacerbate them.[24]

Digitally dominant languages are often described in the AI context as High-Resource Languages, a term referring to the richness of data resources linked to them.[25] They form a very small club, though. Only a few tens of languages are considered High-Resource Languages, a fraction of the thousands of existing languages.[26] The remaining are, unfortunately, Low-Resource Languages.[27] These terms, albeit common, are nonetheless tricky. Their neutral wording glosses over the deep political, societal, and economic disparities among languages, blurring the processes of domination and exclusion that have shaped them. Using these terms, with

---

[21] See *infra* Part II.B.1 for the discussion on this later in this Article.

[22] *Id.*

[23] *See* Ranathunga & de Silva, *supra* note 13, at 823; Wu et al., *supra* note 19; Alexander H.E. Morawa, *Minority Languages and Public Administration: A Comment on Issues Raised in Diergaardt et al. v. Namibia* 1-26, Eur. Ctr. for Minority Issues, Working Paper No. 16, 2002 (explaining that most of the world's languages "have been and are still ignored in the aspect of language technologies.").

[24] *See generally id.* (referring to the "digital divide" between high-resource and low-resource languages).

[25] *Id.*

[26] Emily M. Bender, *The #BenderRule: On Naming the Languages We Study and Why It Matters,* THE GRADIENT (Sept. 14, 2019), https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/ [https://perma.cc/6RR2-VWLB].

[27] Ranathunga & de Silva, *supra* note 13, at 823. Low-resource languages are sometimes called "low-density languages," "under-resource languages," or "low data languages." *See* L. Besacier et al., *Automatic Speech Recognition for Under-Resourced Languages: A Survey,* 56 SPEECH COMMC'N 85, 87 (2014). Some researchers also refer to a mid-resource category, see generally: Pedro Javier Ortiz Suárez et al., *A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages*, *in* PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 1703 (Dan Jurafsky & Joyce Chai et al., eds. 2020).

https://doi.org/10.59015/wilj.SCTW4297

their neutral wording, might also encourage a passive view of the existing linguistic gaps, rather than prompting action to reduce them. Indeed, processes of sidelining and dominating in the linguistic realms are often subtle, "invisible to the casual observer,"[28] and sometimes seamlessly addressed as a modern, almost inevitable phenomenon.[29] I will, therefore, use the terms *Digitally Marginalized Languages* (DMLs) and *Digitally Dominant Languages* (DDLs), to better capture the nature of these linguistic categories and the broader power dynamics surrounding them. Using these terms also sharpens the understanding that linguistic digital inequalities are not a fated phenomenon and that they can and should be mitigated.

Several techno-social factors drive the disparities between DMLs and DDLs in LLMs. In this Article, I discuss two main categories of such causes: (1) training datasets and training processes, and (2) design and evaluation choices and constraints.[30] Regarding the first category, one key challenge is that most DMLs do not have enough available training data to power a language-specific LLM.[31] Most of these languages are also absent from multilingual models' pretraining (unsupervised) processes, let alone fine-tuning (supervised) processes that require labeled data.[32] Often, these multilingual models' training sets only encompass a few tens of languages or around one hundred languages in several other cases.[33] Moreover, DMLs that *are* represented in the training data of such models only constitute a small fraction of the entire dataset used in the pretraining stage.[34] The limited scope of DML data in multilingual models not only hinders the models' linguistic performance in these languages, but may

---

[28] Eriksen, *supra* note 5, at 313.

[29] *See id.* at 313-314.

[30] See *infra* Part II.B.2 for the discussion on this later in this Article.

[31] *See*, *e.g.*, Monojit Choudhury, *Generative AI has a Language Problem*, 7 NAT. HUM. BEHAV. 1802 (2023).

[32] Sumanth Doddapaneni et al., *Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages*, *in* 1 PROCEEDINGS OF THE 61ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 12402, 12402-03 (Anna Rogers & Jordan Boyd-Graber et al., eds. 2023). Pretraining is the process through which LLMs are exposed to large corpora of unlabeled data. This is often followed by fine-tuning, in which the model is further trained on a specific language or task using labeled data. See *infra* Part II.A for the discussion on this later in this Article.

[33] See *infra* Part II.B.2. for the discussion on this later in this Article.

[34] Doddapaneni et al., *supra* note 32, at 12402-03.

https://doi.org/10.59015/wilj.SCTW4297

also result in LLMs that are culturally skewed toward very confined norms, knowledge, and values.[35]

The other category of challenges relates to design choices and constraints that may discriminate against DML speakers and limit their ability to engage with LLMs. It includes tokenization processes that favor Latin and Cyrillic languages and filtering practices that silence vulnerable groups, including speakers of so-called African American-aligned English and Hispanic-aligned English. This category also relates to the evaluation sets picked by LLM developers, which are often in DDLs. Finally, it concerns the worrying lack of benchmarks in DMLs and the resulting translation of DDL benchmarks to DMLs, which may bake in mistakes, along with linguistic and cultural misalignments.[36]

To capture the nature and implications of the digital sidelining of DMLs, I draw on Nancy Fraser's "Parity of Participation" theoretical framework. Fraser perceives "Parity of Participation" as the most general meaning of justice.[37] Such parity, she explained, requires "social arrangements that permit all to participate as peers in social life."[38] These arrangements can only be achieved if

> all the relevant subjects have no entrenched social obstacles that in a structural way prevent them from participation in terms of parity or equality—whether this is participation in formal and informal political and public spheres, institutions, life, in civil society, in the life of associations, in family life, in labor markets, in fact in any and all of the major institutional arenas that are important in society.[39]

According to Fraser, participation builds on three layers: distribution, recognition, and representation.[40] As I later show, all three are denied to DML speakers. First, maldistribution exists, since the current economic structure systematically excludes DML speakers from a fair share of the resources, opportunities, and freedoms introduced by LLMs.

---

[35] *See* Mehrnaz Siavoshi, *The Importance of Natural Language Processing for Non-English Languages*, MEDIUM (Sept. 21, 2020), https://towardsdatascience.com/the-importance-of-natural-language-processing-for-non-english-languages-ada463697b9d [https://perma.cc/8N63-F7SB] (discussing the consequences of the limited scope of DML data). For discussion on how this may skew cultural perception, see also *infra* Part II.B.2.

[36] *See* Siavoshi, *supra* note 35.

[37] NANCY FRASER, SCALES OF JUSTICE: REIMAGINING POLITICAL SPACE IN A GLOBALIZING WORLD 16 (2009).

[38] *Id.*

[39] Amrita Chhachhi, *Nancy Fraser Interviewed by Amrita Chhachhi*, 42 DEV. & CHANGE 297, 303 (2011).

[40] FRASER, *supra* note 37, at 16–18.

https://doi.org/10.59015/wilj.SCTW4297

Moreover, maldistribution in the NLP and LLM contexts exposes DML speakers to erroneous sanctions applied by both governments and digital platforms.[41] Second, misrecognition—the status of a group as undeserving of an equal place at the table, according to Fraser—also occurs in linguistic digital contexts. The institutionalized devaluation of DMLs' cultural worth has been reflected in pre-digitalization colonialism and linguistic oppression and is embodied in the frequent absence of DMLs from the entire chain of LLM development and assessment.[42] Third, DML speakers are often unrepresented in decision-making processes that shape the linguistic governance landscape. This is reflected, inter alia, in the procedures by which certain new AI regulatory tools were adopted and in the circle of actors allowed within such decision-making processes.[43]

Drawing on the linguistic participation gaps unveiled through the application of Fraser's framework and the techno-social disparities described, I outline a path forward spanning technological, societal, and regulatory considerations, aiming towards a more just digital linguistic landscape.

In tandem with the application of Fraser's work and the techno-social exploration, international law requirements—particularly the United Nations' Guiding Principles on Business and Human Rights (UNGPs)—will inform my suggestions for creating a more inclusive digital linguistic future.[44] Such international law requirements concern states' obligations to facilitate the right to equality and private companies' responsibilities to respect this right, and support the grounds for expecting these stakeholders to actively mitigate the participation barriers faced by DML speakers.

Tackling these digital linguistic gaps is time-sensitive. In this regard, Siavoshi noted:

> The fact is that supported systems continue to thrive while it is challenging to introduce new aspects to a deeply ingrained program. This means that as NLP continues to develop without bringing in a

---

[41] *See* Andrew Warner, *Machine Translation Is No Substitute for Humans when It Comes to Law Enforcement*, MULTILINGUAL (Nov. 21, 2022), https://multilingual.com/david-utrilla-expert-witness/ [https://perma.cc/H59H-6U2K]; *see also* Carey L. Biron, *AI's 'Insane' Translation Mistakes Endanger US Asylum Cases*, CONTEXT (Sept. 18, 2023), https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases [https://perma.cc/895H-LU6Y]. For further discussion, see *infra* Parts III.A., B.1.

[42] *See infra* Part III.A, B.2.

[43] *See infra* Parts III.A, and B.3

[44] *See infra* Part V.

https://doi.org/10.59015/wilj.SCTW4297

> diverse range of languages, it will be more challenging to incorporate them in the future, endangering the global variety of languages.[45]

We should, therefore, act now.

This Article proceeds as follows: The first part addresses linguistic disparities in digital avenues and situates these gaps within a broader contextual frame of oppression and control. The second part explores NLP and LLM technologies as an influential case study of digitalization. It describes these technologies' development and provides a detailed account of the linguistic opportunities and limitations they introduce. The third part analyzes contemporary digital linguistic disparities through the lens of Nancy Fraser's "Parity of Participation" theoretical framework, covering the maldistribution of resources and DML speakers' misrecognition and misrepresentation. The fourth part discusses the legal duties and responsibilities of governments and private companies to facilitate DML speakers' digital linguistic participation. The fifth part integrates the technological, fairness, and legal perspectives, offering a way forward and outlining the steps that governments and private AI companies should take to foster diversity in digital linguistic avenues.

I. WHOSE VILLAGE IS THIS?

Digitalization's far-reaching opportunities are not equally allocated among speakers of different languages. While DDL speakers, and most prominently, English speakers, often have full access to myriad digital spaces and their benefits, speakers of DMLs—the lion's share of the world's languages—are either entirely excluded from these possibilities or provided with limited access to them. Indeed, as noted in a joint report published by The Centre for Internet and Society, Oxford Internet Institute, and Whose Knowledge?, "[t]he internet is not yet (and sadly, nowhere near) as multilingual as we are in real life."[46] In this part, I will address some pressing manifestations of the digital linguistic gaps. I will also discuss the long-standing power, oppression, and control dynamics surrounding contemporary linguistic disparities.[47]

---

[45] Siavoshi, *supra* note 35.

[46] *Id.*

[47] *See generally* Sophie Lythreatis et al., *The Digital Divide: A Review and Future Research Agenda*, TECH. FORECASTING & SOC. CHANGE, Feb. 2022, at 1, 1 (explaining how digital gaps are not limited to linguistic concerns, the term "digital divide" is often used to generally describe the unjust allocation of digitalization benefits).

https://doi.org/10.59015/wilj.SCTW4297

A. LINGUISTIC DISPARITIES AND DIGITALIZATION

What predicaments do speakers of DMLs face when attempting to participate in digital avenues while using their first language[48] (rather than switching to a DDL)?[49]

The first category of such difficulties goes back to basic connectivity requirements. While unprecedented digital breakthroughs transform the human experience, over a third of the world's population is not connected to the internet.[50] Among those unconnected, many are from Asia and Africa.[51] Speakers of the languages spoken in such unconnected areas are thus entirely excluded from participating in everything digital and the extensive opportunities it grants.[52]

Another category of challenges regards equipment. Physical keyboards in DMLs are often hard to find. Some handle this shortage by attaching DML letters to the devices. However, this method is not always feasible. For example, it may be challenging to attach DML letters to small devices, such as phones with physical keyboards,[53] which are prevalent in many African countries, among other places.[54] Visual keyboards can

---

[48] *See generally* J.A. Burn et al., *'I Study Long, Long Time in My Language, so I Never Forget It': Reading and First Language Maintenance*, 25 INTERCULTURAL EDUC. 377, 378 (2014), https://www.tandfonline.com/doi/full/10.1080/14675986.2014.967974?needAccess=true#d1e161 [https://perma.cc/G3AQ-GCN3] (explaining how "First language" has been defined using different terms, including "mother tongue," "primary language," or "native language," all of which refer to languages acquired during one's early childhood. However, this is not necessarily the language one identifies with, uses the most, or is fluent in).

[49] *See generally* Katy E. Pearce & Ronald E. Rice, *The Language Divide—The Persistence of English Proficiency as a Gateway to the Internet: The Cases of Armenia, Azerbaijan, and Georgia*, 8 INT'L J. OF COMMC'N 2834, 2839 (2014), https://ijoc.org/index.php/ijoc/article/view/2075 [https://perma.cc/CUF8-4LNB] (noting that DML speakers often *cannot* switch to a DDL as a gate for digital participation, due to lack or limited proficiency in that language).

[50] *ITU Report: One-third of the Global Population Remains Unconnected*, DIGWATCH (Sept. 14, 2023), https://dig.watch/updates/itu-report-one-third-of-the-global-population-remains-unconnected#:~:text=The%20latest%20numbers%20from%20ITU,the%20global%20population%20is%20unconnected [https://perma.cc/9BFV-4C72].

[51] *Countries with the Highest Number of People Not Connected to the Internet as of October 2024*, STATISTA, https://www.statista.com/statistics/1155552/countries-highest-number-lacking-internet/ [https://perma.cc/7GYF-DD3L] (last visited Jan. 30, 2025); *see also Internet Speeds by Country*, WORLD POPULATION REV., https://worldpopulationreview.com/country-rankings/internet-speeds-by-country [https://perma.cc/X5DR-VJ53] (last visited Jan. 30, 2025) (showing differences also exist concerning the internet's quality).

[52] *See ITU Report*, *supra* note 50.

[53] *Id.*

[54] Laura Silver & Courtney Johnson, *Internet Connectivity Seen as Having Positive Impact on Life in Sub-Saharan Africa*, PEW RSCH. CTR. 12 (Oct. 9, 2018),

https://doi.org/10.59015/wilj.SCTW4297

assist, in some cases. For some products, visual keyboards are available in hundreds of languages. While valuable, this solution only serves a limited segment of the thousands of languages that exist.[55]

This leads to the third category, technological language support. Many applications and platforms only support a small number of languages (around ten to thirty).[56] Waze, for instance, is available in only twenty-seven languages.[57] Other services, sites, and apps cater to a wider list of languages, but still only cover a small fraction of the world's languages. WhatsApp, for instance, is currently available in about forty to sixty interface languages, depending on the operating system.[58] Google Maps offers more than seventy languages, Google Search about one hundred fifty, Facebook more than one hundred, YouTube more than seventy, and X (previously known as Twitter) more than forty-five.[59] Wikipedia leads the list with more than three hundred interface languages,[60] but this figure still relates to only a small part of the world's languages.

The fourth category of obstacles that DML speakers might encounter concerns the content available to them. Many languages have a limited or no written system, including sign languages and other languages that are not script-based. This dramatically limits the availability of these languages' content in digital domains.[61] Nonetheless, many script-based languages also face significant content challenges. English almost

---

https://www.pewresearch.org/global/wp-content/uploads/sites/2/2018/10/Pew-Research-Center_Technology-use-in-Sub-Saharan-Africa_2018-10-09.pdf [https://perma.cc/2WSN-S5L8].

55 *See* Daan van Esch et al., *Writing Across the World's Languages: Deep Internationalization for Gboard, the Google Keyboard*, ARXIV (Dec. 3, 2019), http://arxiv.org/abs/1912.01218 [https://perma.cc/RKV5-ZHYN].

56 Martin Dittus & Mark Graham, *A Platform Survey: Interface Language Support by Widely-Used Websites and Mobile Apps*, STATE OF THE INTERNET'S LANGUAGES REP., https://internetlanguages.org/en/numbers/a-platform-survey/ [https://perma.cc/32W9-8G65] (last visited Jan. 30, 2025).

57 *Countries and Languages*, WAZEOPEDIA, https://www.waze.com/wiki/USA/Countries_and_Languages [https://perma.cc/SKU3-LFYH] (last visited Jan. 30, 2025).

58 *How to Change WhatsApp's Language*, WHATSAPP, https://faq.whatsapp.com/779773243128935 [https://perma.cc/XKM3-XU99] (last visited Jan. 30, 2025).

59 Dittus & Graham, *supra* note 56.

60 Martin Dittus & Mark Graham, *The Language Geography of Wikipedia*, STATE OF THE INTERNET'S LANGUAGES REP., https://internetlanguages.org/en/numbers/wikipedia-language-geography/ [https://perma.cc/R9VP-AXV9] (last visited Jan. 30, 2025); *see also infra* Part I.A. (discussing gaps in Wikipedia's coverage across languages).

61 *Id.*

https://doi.org/10.59015/wilj.SCTW4297

exclusively dominated the internet in its early years,[62] and still leads the list of online content with more than 50 percent of the global web content.[63] The following nine leading languages lag significantly behind English, each with about 2–6 percent of the global web content.[64] Still, together with English, this small group of languages accounts for around 85 percent of the online content, highlighting the profound content inequalities among languages.[65]

Martin Dittus and Mark Grahm examined such difficulties in the context of Google Maps. They found that around half of the sites they had covered in their research appeared in English-language searches. However,

> only 20-25 percent of these places were included in the results for French, Spanish, Russian and Portuguese searches, and only 10-15 percent in the search results for Indonesian, Arabic, and Mandarin Chinese. By contrast, speakers of Hindi were shown less than five percent of the global map, and speakers of Bengali less than one percent of the global map.[66]

Google Maps in English, they more specifically explained, covers the entire world (though the coverage is denser in the Global North). Conversely, Google Maps in Hindi, the third most spoken language in the world, is confined to South Asia, particularly India and Bangladesh.[67] Dittus and Graham stressed that such disparities exist not only in global online coverage of information but also on the local level. Google Maps coverage of Kolkata, for instance, is better in English than in Hindi and Bengali, though the three languages are spoken in that city.[68] The researchers unveiled the same content gaps across languages on Wikipedia.[69] They found that "Wikipedia's language editions vary widely in scale—both in terms of number of articles, but also in terms of the size of their editor communities."[70] English Wikipedia, Dittus and Graham

---

[62] Daniel Pimienta et al., *Twelve Years of Measuring Linguistic Diversity in the Internet: Balance and Perspectives*, UNESCO PUBL'NS FOR THE WORLD SUMMIT ON THE INFO. SOC'Y 1 (2009), https://unesdoc.unesco.org/ark:/48223/pf0000187016 [https://perma.cc/77KQ-CXR3].

[63] Petrosyan, *supra* note 8.

[64] *Id.* (these languages are, by order, Spanish, German, Russian, Japanese, French, Portuguese, Italian, Turkish, Dutch/Flemish).

[65] Petrosyan, *supra* note 8.

[66] *Id.*

[67] *See Summary Report supra* note 6; Dittus & Graham, *supra* note 56.

[68] Dittus & Graham, *supra* note 56.

[69] *See Summary Report*, *supra* note 6.

[70] Dittus & Graham, *supra* note 60.

https://doi.org/10.59015/wilj.SCTW4297

observed, leads the list, with about six million articles and forty million contributors. Following are Wikipedia's Spanish, German, and French editions, with around two million articles and four to six million contributors.[71] Other Wikipedia editions remain far behind, with only "a small fraction of the content that is found in English Wikipedia."[72]

Albeit significant, the gaps between DMLs and DDLs are not confined to matters of the scope or the density of the content. Instead, such gaps also apply to the voices, narratives, and perspectives, visible and represented, within such content. When DML speakers go online, they may find the existing informational resources do not adequately represent their collective histories, lived experiences, values, or needs. This stems from the fact that it is DDL speakers who often tell DML speakers' stories when creating content.[73] This concern is supported by the aforementioned Wikipedia research, which indicated that most of the Wikipedia content describing countries in the Global South was written in a foreign language, rather than the local one.[74]

Online content may, thus, miss out on appropriately representing certain countries and the communities they populate. Navigli et al. noted in this regard: "different people speak not only different languages but also embody different cultures, histories, and traditions; therefore, they value different topics with varying degrees of importance."[75] Moreover, using one language to create knowledge and describe developments and facts relating to communities of different languages may also encapsulate bias or reinforce existing societal and political asymmetries and hierarchies.[76] Mary Talbot et al. explained that language can construct such perceptions through the choice of issues that will be discussed, the context they will

---

[71] *Id.*

[72] *Id.* ("[O]nly around 20 language editions have more than one million articles, and only 70 have more than 100,000 articles.").

[73] *See, e.g.,* Jon C. Stott, *Native Tales and Traditions in Books for Children*, 16 AM. INDIAN Q. 373, 374 (1992); Judy Iseke-Barnes, *Living and Writing Indigenous Spiritual Resistance*, 24 J. OF INTERCULTURAL STUD. 211, [pincite] (2003); Michael Gawenda, *The Age's Truth: Indigenous Stories Told by White Writers*, THE AGE (Oct. 18, 2021), https://www.theage.com.au/national/victoria/the-age-s-truth-indigenous-stories-told-by-white-writers-20210802-p58f71.html [https://perma.cc/U645-4P9F].

[74] Dittus & Graham, *supra* note 60.

[75] Roberto Navigli et al., *Biases in Large Language Models: Origins, Inventory, and Discussion*, ASS'N COMPUT. MACH. J. OF DATA & INFO. QUALITY, June 2023, at 1, 7.

[76] *See* DILIP CHAVAN, LANGUAGE POLITICS UNDER COLONIALISM: CASTE, CLASS, AND LANGUAGE PEDAGOGY IN WESTERN INDIA (Cambridge Scholars Publ'g ed., 2013); *see also* Stacey Lee, *Dear Non-Asian Writer*, HYPHEN (Feb. 11, 2016), https://hyphenmagazine.com/blog/2016/02/dear-non-asian-writer [https://perma.cc/WQX5-XR96].

https://doi.org/10.59015/wilj.SCTW4297

be embedded in, and even the vocabulary that will be used. For instance, minority groups are often labeled with certain words that the hegemonic white majority does not tend to use when self-referring, thereby further "othering" them.[77] In their book, "Language and Power in the Modern World," Mary Talbot et al. stated, in this regard:

> we would not refer to the Women's Institute as an "ethnic group," to fish and chips as "ethnic food," or to a bowler hat as "traditional ethnic headgear." The expression "ethnic riot" would never be used in the press to describe civil disturbance caused by white middle-class people in the home counties, though the white middle-class of the home counties are just as much an ethnic group as the black community in Brixton, or any other. The white middle-class unselfconsciously occupies the neutral, "non-ethnic" centre.[78]

Due to the nuanced messages and meanings embodied in the language itself, translation—while invaluable—often fails to substitute for original content created in DMLs.[79] For instance, Kelsey Begaye, the Navajo Nation president, responded to a state initiative to hold English-only curriculum in American state schools, saying:

> The Navajo Way of Life is based on the Navajo language. By tradition, the history of our people and the stories of our people are handed down from one generation to the next through oral communication. Naturally, the true essence and meanings for many Navajo stories, traditions and customs cannot be fully transmitted, understood or communicated as told through non-Navajo languages.[80]

Lastly, in cases where DML speakers created content, certain challenges emerged from within. The authors of such content were sometimes members of the more dominant, strong subgroups among that given DML community, resulting in the intentional or unintentional silencing of vulnerable voices within that community, including those of LGBTQ+ persons, women, and people with disabilities. Furthermore, this practice sometimes leads to the generation and spread of content that jeopardizes weak sectors within that community.[81]

---

[77] MARY TALBOT ET AL., LANGUAGE AND POWER IN THE MODERN WORLD 15, 48 (2003).

[78] *Id.*

[79] Alicia Shepard-Vega, *The Consequences of Meta's Multilingual Content Moderation Strategies*, DIGWATCH, (Sept. 1, 2023), https://dig.watch/updates/the-consequences-of-metas-multilingual-content-moderation-strategies [https://perma.cc/5DB3-WK6J].

[80] Kelsey Begaye, *Guest Editorials*, NAVAJO-HOPI OBSERVER (Sept. 19, 2000), https://www.nhonews.com/opinion/guest-editorials/article_8c10ba17-2aab-5d5a-913b-6a75ad4c2714.html [https://perma.cc/XD5B-8AEN].

[81] *See Summary Report*, *supra* note 6; Darmawan, *supra* note 12.

https://doi.org/10.59015/wilj.SCTW4297

B. LANGUAGES, OPPRESSION, AND DIGITALIZATION

1. The Enduring Roots of Linguistic Oppression

Of course, linguistic inequalities existed long before digitalization became ingrained in our lives. The history of languages is interlocked with discrimination, oppression, and domination. It was shaped by impactful, at times violent, processes, some of which occurred at the national level, and some were powered by colonialism and imperialism.[82]

The establishment of nation-states and the rise of nationalism encouraged a hierarchy of languages, much of which is still evident today. Within this new order, a small group of languages enjoyed official recognition,[83] while attempts were often made to limit or even eradicate minority languages.[84] Formal and informal policies and practices around languages were also used to identify, create, and sustain class structures within a given society based on the dialects, accents,[85] and vocabulary in use.[86] These processes often included the banning of indigenous languages and various sanctions, some violent, for violating the ban.[87]

Colonialism and imperialism also play a dramatic part in today's global language settings.[88] This is reflected, for instance, in the fact that five of the ten most spoken languages in the world (English, Spanish,

---

[82] *See e.g.,* Eriksen, *supra* note 5; Gerald Roche, *Articulating Language Oppression: Colonialism, Coloniality and the Erasure of Tibet's Minority Languages*, 53 PATTERNS OF PREJUDICE 487, 488 (2019) (engaging with the work of Alice Taff and others, Roche notes that "Language oppression," is a "form of domination that is coherent with other forms of oppression along the lines of 'race', nation, colour and ethnicity"); *see* Alice Taff et al., *Indigenous Language Use Impacts Wellness, in* THE OXFORD HANDBOOK OF ENDANGERED LANGUAGES 862, 863 (Kenneth L. Rehg & Lyle Campbell eds., 2018).

[83] Eriksen, *supra* note 5, at 313.

[84] *Id.* at 320.

[85] RICHARD BAUMAN & CHARLES L. BRIGGS, VOICES OF MODERNITY: LANGUAGE IDEOLOGIES AND THE POLITICS OF INEQUALITY 7 (Cambridge Univ. Press 2003) (addressing the impact of educators "who make non-standard dialects into markers of irrationality, ignorance, school failure, and suitability for dead-end service jobs"); *see also* TALBOT ET AL., *supra* note 77, at 12 (discussing the canonical status of "Received Pronunciation" in British broadcasting).

[86] *See, e.g.*, BAUMAN & BRIGGS, *supra* note 85, at 42.

[87] Lenore A. Grenoble, *Language ecology and endangerment*, *in* THE CAMBRIDGE HANDBOOK OF ENDANGERED LANGUAGES 27, 32 (Julia Sallabank & Peter K. Austin eds., 2012); *see also* JANE GRIFFITH, WORDS HAVE A PAST 65 (Univ. Toronto Press 2019).

[88] *Europe Population (Live)*, WORLDOMETER, https://www.worldometers.info/world-population/europe-population/ [https://perma.cc/2URD-RAQ7] (last visited Nov. 29, 2023); *see also* Abram de Swaan, *The Emergent World Language System: An Introduction*, 14 INT'L POL. SCI. REV. / REVUE INTERNATIONALE DE SCI. POLITIQUE 219, 220 (1993); BRUCE MANNHEIM, THE LANGUAGE OF THE INKA SINCE THE EUROPEAN INVASION 80 (Univ. Texas Press 2013).

https://doi.org/10.59015/wilj.SCTW4297

French, Portuguese, and Russian) are European in origin,[89] even though Europe's population only constitutes about 9 percent of the world's population,[90] and that this continent has the fewest languages in the world compared to other continents.[91] In addition to the creation of nation-states and processes of nationalism and colonialism, modernization and urbanization have also greatly influenced many languages. In the modern world—characterized by modern education, modern media and communication, and modern professions—there is often a tendency to smooth out cultural differences, including those of a linguistic nature.[92] This further strengthens the already dominant languages and creates a subtle system that penalizes the use of more vulnerable languages. Some warn that this trend "encourages a mass of inferiority complexes and the eventual abandonment of maternal languages among minorities."[93]

All these processes contributed to the weakening of minority and local languages, and even to their endangerment.[94] Currently, over 40 percent of the world's languages are endangered.[95]

Indeed, languages were often the subject of uneven power allocation or the means to it. The linguist Tove Kuttab-Kangas coined the term "linguicism" to capture language discrimination. She defined it as: "ideologies, structures and practices which are used to legitimate, effectuate, regulate and reproduce an unequal division of power and

---

[89] These 10 languages are, by order of their speaker base: English, Mandarin Chinese, Hindi, Spanish, French, Arabic, Bengali, Portuguese, Russian, and Urdu. *What are the top 200 most spoken languages?*, ETHNOLOGUE, https://www.ethnologue.com/insights/ethnologue200/ [https://perma.cc/MM2X-NZUX] (last visited Aug. 1, 2024); *see also* Roche, *supra* note 82, at 488–490 (describing the relationship between language erasure and imperialism).

[90] *Population by Continent 2024*, WORLD POPULATION REV., https://worldpopulationreview.com/continents [https://perma.cc/9L2E-CW2W] (last visited Jun. 30, 2024).

[91] *Summary Report*, *supra* note 6, at 7; *see also* Grenoble, *supra* note 87, at 28 ("Nearly a third of all languages are spoken in Asia, and 30% are spoken in Africa. Only 3.5% are spoken in Europe, and under 15% are spoken in North and South America combined.").

[92] *See* Eriksen, *supra* note 5, at 316–318.

[93] *Id.* at 318; *see also* Grenoble, *supra* note 87, at 32.

[94] Robert Phillipson & Tove Skutnabb-Kangas, *Linguistic Imperialism and Endangered Languages*, *in* THE HANDBOOK OF BILINGUALISM AND MULTILINGUALISM 495, 495 (Tej K. Bhatia & William C. Ritchie eds., 2012) (in the context of colonialism and imperialism); *How many languages are endangered?*, ETHNOLOGUE, https://www.ethnologue.com/insights/how-many-languages-endangered/ [https://perma.cc/S6FU-6GG8] (last visited Jul. 22, 2024) ("Institutional languages are least likely to become endangered – they have been adopted by governments, schools, mass media, and more.").

[95] *How many languages are endangered?*, *supra* note 94.

https://doi.org/10.59015/wilj.SCTW4297

resources…between groups which are defined on the basis of language."[96] Kuttab-Kangas further explained that "[m]ost practices where people get unequal access to power and both material and immaterial resources, based on their language/s, reflect linguicism."[97] Lionel Wee's writings reflect this same notion, stating that:

> There are many situations where language is seen to play an invidious role in the perpetuation of social inequity. Such situations, broadly speaking, involve individuals or groups being denied access to social and economic goods, or even a sense of dignity and pride in their own identities, simply by virtue of the language that they happen to speak.[98]

Heller and McElhinny explored how language "is used to make boundaries that help produce, reproduce, or contest" unequal distribution of resources.[99] They have also examined how capitalism and colonialism have supported "particular ways of mobilizing language in the production of inequality and social differences that legitimize it."[100]

Albeit impactful, the execution of linguistic discriminative power may be very subtle and hard to detect. According to Thomas Eriksen, the forms of linguistic oppression of minorities are "not usually of a physical and spectacular kind. On the contrary, they are often invisible to the casual observer, and they are sometimes not even articulated as forms of oppression."[101] Therefore, he further noted, oppressive processes were not investigated as such. Instead, "they have been described and analyzed as processes of modernization or—more generally—of social change, as minority strategies, cultural homogenization, or cultural conflict."[102]

2. Digitalization and Linguistic Exclusion

How does digitalization fit into this environment of intense domination and exclusion? While the digital age has afforded many opportunities for DML speakers, some of which will be discussed

---

[96] Tove Skutnabb-Kangas, *Linguicism*, *in* THE ENCYCLOPEDIA OF APPLIED LINGUISTICS 1, 1 (Carol A. Chapelle ed., 1st ed. 2015).

[97] *Id.* at 2.

[98] LIONEL WEE, LANGUAGE WITHOUT RIGHTS 3 (Oxford Univ. Press 2011).

[99] MONICA HELLER & BONNIE MCELHINNY, LANGUAGE, CAPITALISM, COLONIALISM: TOWARD A CRITICAL HISTORY 3 (Univ. Toronto Press 2017).

[100] *Id*. at 3–4 (noting how language has been used to both entrench and combat inequalities in colonialist and capitalist histories).

[101] Eriksen, *supra* note 5, at 313.

[102] *Id.*

https://doi.org/10.59015/wilj.SCTW4297

below,[103] these opportunities are a watered-down version of the benefits offered to speakers of DDLs.[104] Rather than rejoicing with the limited advancements for DMLs, we should see things for what they are: digitalization is developed and deployed in a fashion that perpetuates hierarchical societal and political power structures and further sidelines already marginalized languages. It is important to emphasize that minority languages or those spoken by small communities are not the only ones affected by digitalization inequalities. The impact of these inequalities is much wider, covering some languages with the largest speaker bases worldwide, including Hindi, Bengali, and Urdu.[105]

It seems, thus, that the world is hardly "one small village" for those on the wrong side of the linguistic landscape. Instead, this romantic village, if it exists, only sustains a fraction of the global languages, and firstly, English.[106] These gaps are worryingly reflected in advanced NLP and LLM technologies, the contemporary transformative representations of digitalization.

In the following Parts, I will discuss how these technologies embody and reinforce the linguistic sidelining processes reflected in previous digital developments. First, I will explore key milestones in NLP and LLM development. After setting this technological ground, I will examine the benefits and challenges these technologies introduce for the global linguistic landscape.

II. THE CASE OF NLP AND LLMS

The following sections will anchor the discussion on digital linguistic disparities in one field of advanced AI technologies: NLP-driven technologies, and more specifically, LLMs. Exploring these technologies is beneficial as their relation to language is particularly close. Moreover, they are emerging as digital architectures that revolutionize the human experience in numerous ways and are expected to be further developed and incorporated into future AI infrastructures.[107]

---

[103] *See infra* Part II.B.1.

[104] *See supra* Part I; *infra* Part II.B.2.

[105] Besacier et al., *supra* note 27, at 87; *Summary Report*, *supra* note 6, at 8.

[106] Petrosyan, *supra* note 8.

[107] See discussion *infra* Part II.

https://doi.org/10.59015/wilj.SCTW4297

A. COMPUTATIONAL LINGUISTICS—NATURE AND DEVELOPMENT

1. NLP and Linguistics

NLP, also known as computational linguistics, is a branch of computer science, and more specifically, of AI. It concerns the computational methods for acquiring, understanding, and producing languages.[108] NLP builds on a set of fields and disciplines, including linguistics, machine learning, deep learning, and statistical models.[109] It serves as the technological foundation for various applications such as machine translation, chatbots, content moderation, personal assistance, spellcheck, email spam detection, information extraction, and content summarization.[110]

Linguistics plays an important role in the development of NLP.[111] It covers different subfields,[112] among which are syntax (sentences' formation and structure);[113] semantic (sentences' meaning);[114] discourse (analysis of text, beyond a single sentence); phonology (the sound system); and pragmatics (contextual meaning of content).[115]

---

[108] Julia Hirschberg & Christopher D. Manning, *Advances in Natural Language Processing*, 349 SCI. 261, 261 (2015); *What is Natural Language Processing?*, IBM, https://www.ibm.com/topics/natural-language-processing [https://perma.cc/A5EP-LNHL] (last visited Dec. 6, 2023).

[109] Eda Kavlakoglu, *NLP vs. NLU vs. NLG: The Differences Between Three Natural Language Processing Concepts*, IBM (Nov. 12, 2020), https://www.ibm.com/think/topics/nlp-vs-nlu-vs-nlg [https://perma.cc/49SM-29P4].

[110] Khurana et al., *supra* note 4, at 3724; *Lost in Translation: Large Language Models in Non-English Content Analysis*, CTR. FOR DEMOCRACY & TECH. (May 23, 2023), https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ [https://perma.cc/CVD4-SUUU].

[111] Khurana et al., *supra* note 4, at 3715.

[112] *Id.*

[113] Tatwadarshi P. Nagarhalli et al., *Impact of Machine Learning in Natural Language Processing: A Review,* 2021 THIRD INT'L CONF. ON INTELLIGENT TECHS. & VIRTUAL MOBILE NETWORKS (ICICV) 1529, 1531 (describing how syntax may encompass the elements required in a sentence and the order of the words/elements within a sentence).

[114] *Id.* at 1532.

[115] EMILY M. BENDER, LINGUISTIC FUNDAMENTALS FOR NATURAL LANGUAGE PROCESSING 1 (Graeme Hirst ed., 2013). Unlike semantics, pragmatics does not concern the literal meaning of words, but "what speaker implies and what listener infers." Khurana et al., *supra* note 4, at 3718; *see also* Daniel Hershcovich & Lucia Donatelli, *It's the Meaning That Counts: The State of the Art in NLP and Semantics*, 35 KÜNSTLICHE INTELLIGENZ 255, 257 (2021).

https://doi.org/10.59015/wilj.SCTW4297

Familiarity with these linguistic subfields can inform and enhance the design of NLP-based technologies.[116] Part of Speech tagging, for instance, is an NLP task involving the syntactic labeling of words as nouns, verbs, adjectives, and so on,[117] and Word Sense Disambiguation is a process used to unveil the semantic meaning of a sentence.[118] Some of these tasks are considered "low-level" NLP tasks, while others are "high-level" and require more complex reasoning abilities. Nonetheless, the former are the building blocks of the latter, and both categories are required for high-quality NLP performance.[119]

2. Developments in Advanced NLP Technologies

Since 2010, NLP has been extensively using "neural networks," a technology inspired by the human brain's neural connections.[120] This approach is the backbone of "deep learning," a subbranch of machine learning.[121] Deep learning replaced earlier NLP methods, including certain rule-based[122] and statistical approaches.[123] An even earlier NLP method

---

[116] BENDER, *supra* note 115, at 1 (stressing that such knowledge "can also inform error analysis for NLP systems").

[117] Sujatha Mudadla, *What Is Parts of Speech (POS) Tagging Natural Language Processing?*, MEDIUM (Nov. 9, 2023), https://medium.com/@sujathamudadla1213/what-is-parts-of-speech-pos-tagging-natural-language-processing-in-2b8f4b07b186 [https://perma.cc/75AT-3KXX].

[118] Additional NLP tasks include Reference Resolution (RR), Sentence Splitting, Named Entity Recognition, Event Extraction, Question Answering, and Natural Language Inference (NLI). Nagarhalli et al., *supra* note 113, at 1531-32; Lai et al., *supra* note 15, at 4; *The Stanford Natural Language Inference (SNLI) Corpus*, THE STANFORD NLP GROUP, https://nlp.stanford.edu/projects/snli/ [https://perma.cc/2LXW-3653] (last visited Jul. 4, 2024).

[119] *See, e.g.*, Tommaso Caselli et al., *When It's All Piling up: Investigating Error Propagation in an NLP Pipeline,* 1386 CEUR WORKSHOP PROC. 1, 1 (2015).

[120] *Neural Networks and How They Work in Natural Language Processing*, PANGEANIC (Feb. 23, 2023), https://blog.pangeanic.com/neural-networks-and-how-they-work-in-natural-language-processing [https://perma.cc/4G6K-ACZ8]; *What's the Difference Between Deep Learning and Neural Networks?*, AMAZON WEB SERVS., INC., https://aws.amazon.com/compare/the-difference-between-deep-learning-and-neural-networks/ [https://perma.cc/Z5HV-MQGF] (last visited Dec. 9, 2023).

[121] Tom Young et al., *Recent Trends in Deep Learning Based Natural Language Processing*, IEEE COMPUTATIONAL INTEL. MAG., Aug. 9, 2017, at 1. Machine learning is the subfield of AI that concerns computers' ability to perform tasks for which they were not explicitly programmed, based on examples they were exposed to. Nagarhalli et al., *supra* note 113, at 1530.

[122] Rule-based models focus on grammar, structure, and patterns in the relevant language. It has significant drawbacks, "owing to the variability, ambiguity, and context-dependent interpretation of human languages." Hirschberg & Manning, *supra* note 108, at 261.

[123] *Id.*

https://doi.org/10.59015/wilj.SCTW4297

was word-for-word machine translation, in which English and Russian were the dominant languages in use.[124]

Deep learning has dramatically enhanced NLP capabilities by affording access to complex patterns and insights captured within massive amounts of data.[125] An early breakthrough in deep learning NLP built on a word analysis architecture called "word embeddings" in which words are represented as numerical vectors in one multi-dimensional space. This way, the semantic and pragmatic connections between words, contexts, and analogies are represented by proximity in their vectors.[126] Word embeddings thus introduce advancements in understanding contexts,[127] short forms of writing,[128] idioms, and dual-meaning terms.[129] Word embeddings also allow the identification of word meanings across different languages.[130]

Recent years have witnessed the introduction of LLMs such as ChatGPT (OpenAI), Gemini (Google), and Claude (Anthropic).[131] These models are trained on enormous amounts of data,[132] and demonstrate

---

[124] *See* Khurana et al., *supra* note 4, at 3720.

[125] Saman Razavi, *Deep Learning, Explained: Fundamentals, Explainability, and Bridgeability to Process-Based Modelling*, 144 ENV'T MODELLING & SOFTWARE 105159, 105160 (2021).

[126] Young et al., *supra* note 121, at 2; *see also* Tomas Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality*, 26 ADVANCES IN NEURAL INFO. PROCESSING SYS. 1, 1 (2013). Commonly used deep learning models that leverage the word embedding architecture are CNNs (Convolutional Neural Networks) and RNNs (Recurrent Neural Networks). Young et al., *supra* note 121, at 1.

[127] *See* Nastaran Babanejad et al., *Affective and Contextual Embedding for Sarcasm Detection*, *in* PROC. OF THE 28TH INT'L CONF. ON COMPUTATIONAL LINGUISTICS 225, 225-226 (Donia Scott, Nuria Bel, & Chengqing Zong eds., 2020).

[128] *See* Ahmad Abdulkader et al., *Introducing DeepText: Facebook's Text Understanding Engine*, ENG'G AT META (Jun. 1, 2016), https://engineering.fb.com/2016/06/01/core-infra/introducing-deeptext-facebook-s-text-understanding-engine/ [https://perma.cc/3JVU-8TZ4]; Steven Patterson, *Understanding Deep Text, Facebook's Text Understanding Engine,* NETWORKWORLD (Jun. 01, 2019), https://www.networkworld.com/article/3077998/ understanding-deep-text-facebooks-text-understanding-engine.html [https://perma.cc/V6SA-93FN].

[129] *See Introducing DeepText, supra* note 128.

[130] *Introducing DeepText, supra* note 128. For instance, "Happy birthday" and its parallel Spanish phrase, "Feliz cumpleaños," would probably be proximate in the embedding vector space. *Id.*

[131] *See* Kasneci et al., *supra* note 14, at 1. Many of the existing LLMs leverage the "Transformer" architecture. Relying on the "Self-attention" mechanism, the Transformer allows for the parallel analysis of each word's relationship with others in the text. This speeds up the training process and offers new insights into long-term word dependencies (i.e., involving words that are not necessarily proximate in a given text). *Id.; see also* Ashish Vaswani et al., *Attention is All You Need*, *in* 30 ADVANCES IN NEURAL PROCESSING SYSTEMS 2 (Von Luxburg, U. et al., eds., 2017) (introducing the Transformer architecture).

[132] *How much LLM Training Data Is There, in the Limit?* EDUCATING SILICON (May 9, 2024), https://www.educatingsilicon.com/2024/05/09/how-much-llm-training-data-is-there-in-the-limit/ [https://perma.cc/29F4-XLAU].

https://doi.org/10.59015/wilj.SCTW4297

remarkable abilities in various NLP tasks and speech-processing applications, such as text-to-speech and speech-to-text.[133] Their once-unimagined capabilities span different epistemic fields, including law, healthcare, education, cultural heritage preservation,[134] cognition,[135] scientific research methods, and code writing.[136] They shape our informational landscape by powering search mechanisms[137] and content moderation, among other processes.[138]

A key process in LLM training is transfer learning, a more advanced approach than the "word embeddings" mentioned above.[139] Transfer learning utilizes the knowledge (patterns, relationships, and attributes) captured in pretrained models to learn different, related tasks and languages (the latter is a process called "cross-lingual transfer learning").[140] Pretrained models often leverage unsupervised learning, which relies on unlabeled data throughout the training processes.[141] The pretrained model serves as the starting point for training the new model

---

[133] Yiheng Liu et al., *Summary of ChatGPT-Related Research and Perspective towards the Future of Large Language Models*, META-RADIOLOGY, Aug. 2023, at 1, 1; *see also* Ganesh Joshi *Creating a Voice Recognition System Using NLP Techniques*, MEDIUM (Jul. 17, 2023), https://medium.com/@ganeshchamp39/creating-a-voice-recognition-system-using-nlp-techniques-d724eb395c [https://perma.cc/DL6U-6WHE].

[134] Georgios Trichopoulos, *Large Language Models for Cultural Heritage, in* PROCEEDINGS OF THE 2ND INTERNATIONAL CONFERENCE OF THE ACM GREEK SIGCHI CHAPTER § 6 (CHIGREECE ed., 2023).

[135] Liu et al., *supra* note 133, at 1.

[136] *See id.*

[137] *E.g.,* Pandu Nayak, *Understanding searches better than ever before,* GOOGLE (Oct. 25, 2019), https://blog.google/products/search/search-language-understanding-bert/ [https://perma.cc/R42D-KX56] (explaining BERT is used to facilitate Google search snippets in only 24 countries).

[138] Hakan Inan et al., *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations*, ARXIV 1 at 1 (Dec. 7, 2023), https://arxiv.org/abs/2312.06674 [https://perma.cc/3JFH-ESHM].

[139] Anne Lauscher et al., *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*, ARXIV 1, 2 (2020), http://arxiv.org/abs/2005.00633 [https://perma.cc/HRW2-LYBL]; *but cf.* Tammie Borders & Svitlana Volka, *An Introduction to Word Embeddings and Language Models*, IDAHO NAT'L LAB'Y (Apr. 2021), https://doi.org/10.2172/1773690 [https://perma.cc/URS9-NBFR] (explaining that LLMs, however, often still incorporate Word embeddings as a part of their architecture).

[140] Khwab Kalra, *Transfer Learning and Fine-Tuning*, MEDIUM (Jul. 14, 2023), https://medium.com/@khwabkalra1/transfer-learning-and-fine-tuning-f3db7f7c6ef1 [https://perma.cc/9HYG-H2TV]; *see also* Victor Chaba, *Understanding the Differences: Fine-Tuning vs. Transfer Learning*, DEV CMTY. (Aug. 25, 2023), https://dev.to/luxacademy/understanding-the-differences-fine-tuning-vs-transfer-learning-370 [https://perma.cc/2WQ5-A8ND].

[141] Jiawei Ge et al., *On the Provable Advantage of Unsupervised Pretraining*, ARXIV 1, at 1 (Mar. 2, 2023), https://doi.org/10.48550/arXiv.2303.01566 [https://perma.cc/L2YW-4TFN].

https://doi.org/10.59015/wilj.SCTW4297

and eliminates the need for training from scratch.[142] This reduces time, cost, data, and computational resources that would have been otherwise required.[143]

Transfer learning is often applied in "fine-tuning,"[144] a supervised training methodology where the general-purpose, pretrained model is adapted to enhance the performance on a specific target task or language by leveraging a relatively small set of labeled data.[145] Such labels may include, inter alia, examples of the NLP tasks discussed earlier, to support the models' learning.[146] For machine translation, a subfield of NLP, the labels in the fine-tuning stage will typically include pairs of the source and target languages.[147] High-quality labels are often human-curated and may be expensive and labor-intensive.[148]

Zero-shot learning, an additional, important method that gains much traction in emerging NLP and LLM-based products and research, also relies on transfer learning.[149] However, in zero-shot learning, the pretrained model performs a new task or copes with a new language without the fine-tuning stage, and thus does not necessitate human-curated labels.[150] Zero-shot learning can also apply where fine-tuning has occurred, but the newly acquired NLP task or language was not included in the pretraining dataset.[151]

---

[142] Kalra, *supra* note 140; *see also* Chaba, *supra* note 140.

[143] Kalra, *supra* note 140.

[144] *Id.*; *see also* Chaba, *supra* note 140; *see* Kasneci et al., *supra* note 14, at 2; *see also* Abolfazl Farahani et al., *A Concise Review of Transfer Learning*, ARXIV 1, 1 (Apr. 5, 2021), https://doi.org/10.48550/arXiv.2104.02144 [https://perma.cc/UWC9-DN9X].

[145] Kalra, *supra* note 140; *id.*

[146] *See supra* Part II.A.2.

[147] *See* Rudra Murthy et al., *Addressing Word-Order Divergence in Multilingual Neural Machine Translation for Extremely Low Resource Languages*, *in* 1 PROC. 2019 CONF. N. AM. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS: HUM. LANGUAGE TECH. (LONG & SHORT PAPERS) 3868, 3868 (Jill Burstein et al. eds., 2019).

[148] Aaron Bornstein, *Silver, Gold & Electrum: 3 Data Techniques for Multi-Task Deep Learning,* MEDIUM: TOWARDS DATA SCI. (Aug. 2018), https://towardsdatascience.com/silver-gold-electrum-3-data-techniques-for-multi-task-deep-learning-2655004970a2 [https://perma.cc/E4XH-DGF8] (explaining that human curated labels are sometimes called "Gold labeled data").

[149] Lauscher et al., *supra* note 139, at 1.

[150] Anand Singh, *Leveraging Zero-Shot, Few-Shot, and Transfer Learning: A Comprehensive Guide for Enterprise AI*, MEDIUM (Oct. 1, 2023), https://medium.com/@anand94523/leveraging-zero-shot-few-shot-and-transfer-learning-a-comprehensive-guide-for-enterprise-ai-de64f57e1717 [https://perma.cc/4XCU-M3UM].

[151] Abteen Ebrahimi et al., *AmericasNLI: Evaluating Zero-Shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-Resource Languages*, *in* 1 PROC. 60TH ANN.

Zero-shot learning, driven by DML's data scarcity, has "de facto become the default paradigm for cross-lingual transfer."[152] One-shot learning and few-shot learning are methods involving one or a few examples for each NLP task, respectively.[153] LLMs' training methodology may combine different approaches, including fine-tuning, and zero/one/few-shot learning.[154]

Building on this technological foundation, I will now explore the techno-social advancements and drawbacks brought by NLP and LLMs to DMLs.

B. LLMs Implications for Digitally Marginalized Languages

As mentioned above, LLMs offer humanity groundbreaking advantages, but these benefits are only accessible to a select group of languages.[155] Nonetheless, advancements have been made and invaluable efforts have been directed by various stakeholders at alleviating these gaps.

Exploring this progress provides a more comprehensive account of the digital linguistic ecosystem and highlights desirable directions for further diversification.[156] I will, therefore, begin by examining attempts to broaden the range of languages that can access the benefits of LLMs, as well as their further promise in this regard. I will continue this discussion by mapping and analysing the pressing challenges still faced by DML speakers.

Meeting Ass'n for Computational Linguistics 6279, 6279 (Smaranda Muresan et al., eds., 2022).

[152] Lauscher et al., *supra* note 139, at 1.

[153] Tom B. Brown et al., *Language Models are Few-Shot Learners*, arXiv 1, 4 (May 28, 2020), https://arxiv.org/abs/2005.14165 [https://perma.cc/95ZK-YL7A].

[154] *Id.*; *see also* Zhaofeng Wu et al., *Continued Pretraining for Better Zero- and Few-Shot Promptability*, arXiv (Oct. 21. 2022), https://doi.org/10.48550/arXiv.2210.10258 [https://perma.cc/6EGN-ADGU]; *see infra* Part II.B.2. (explaining that for languages not used in the fine-tuning stage of a certain NLP task, the method typically employed is zero-shot learning or one/few-shot learning, depending on the number of examples in that language).

[155] *See supra* Introduction and *infra* Part II.B.

[156] For the notion of "ecosystem" in the linguistic settings, see *infra* Part III.B.1. It is important to note that the landscape of LLMs is rapidly advancing and evolving. Consequently, some recent developments may not be captured within this Article.

https://doi.org/10.59015/wilj.SCTW4297

1. LLMs' Contribution to the Linguistic Ecosystem

Developments in NLP and LLMs hold the promise to enhance the state of DMLs through two central channels: multilingualism and local, language-specific efforts.

Current multilingual LLMs can perform different tasks across various languages and introduce significant new opportunities for speakers of some DMLs. For instance, these models support machine translation, thus enabling DML speakers to "access information and engage in interlingual dialogue and conversations."[157] These models can also allow DML speakers to interact with administrative bodies that do not offer communication channels in these languages.[158] These DML speakers can now generate content in either their first language or DDLs, access new information in different domains, and use services that were once out of reach.[159]

Multilingual models are often pretrained on multilingual datasets (though English data still constitutes most of these corpora).[160] After the pretraining stage, multilingual language models can leverage transfer learning methodologies through fine-tuning with labeled data or by applying zero-, one-, or few-shot learning approaches, which require few or no examples.[161]

LLMs' architecture and learning approaches leverage the rich linguistic information captured in models pretrained predominantly on DDLs, to train DML models where labeled and unlabeled data is much scarcer.[162] Indeed, multilingual models represent a new generation of models that "significantly boost the performance for NLP tasks in different languages,"[163] to the point where some of these models "have achieved

---

[157] ELIN HAF GRUFFYDD JONES ET AL., NEW TECHNOLOGIES, NEW SOCIAL MEDIA AND THE EUROPEAN CHARTER FOR REGIONAL OR MINORITY LANGUAGES 22 (Jarmo Lainio, ed., 2019).

[158] For use of minority languages, see Morawa, *supra* note 23.

[159] Jones et al., *supra* note 157, at 22.

[160] Lai et al., *supra* note 15, at § 1 ("Similar to other LLMs, ChatGPT is trained on a mix of training data from multiple languages. Although English is the majority, the combination of multilingual data contributes to ChatGPT's abilities to accept inputs and generate responses in different languages, making it accessible and widely adopted by people around the world"); *see also supra* Part II.B.2.

[161] Yuan et al., *How Vocabulary Sharing Facilitates Multilingualism in LLaMA?*, ARXIV § 1–2 (Jun. 3, 2024), https://arxiv.org/html/2311.09071v2 [https://perma.cc/H4QS-QMQB].

[162] *See generally* Telmo Pires et al., *How Multilingual Is Multilingual BERT?*, ARXIV (Jun. 4, 2019), https://doi.org/10.48550/arXiv.1906.01502 [https://perma.cc/F7PQ-G8RX].

[163] Lai et al., *supra* note 15, at § 2.

https://doi.org/10.59015/wilj.SCTW4297

state-of-the-art multilingual performance" in certain NLP tasks.[164] As such, these models can facilitate DML speakers' access to digital avenues and the benefits they embody.[165]

As LLMs improve, so do their multilingual abilities. Multilingual Bert (M-Bert), for instance, was pretrained on Wikipedia datasets in 104 different languages and has introduced significant capabilities.[166] In an evaluation covering sixteen languages, the model demonstrated good cross-lingual generalization through transfer learning.[167] GPT-4 is another example. According to OpenAI, it outperforms its predecessor, GPT-3.5, for most of the languages the company has tested, including DMLs such as Swahili, Latvian, and Welsh.[168]

In a study conducted by Microsoft India,[169] researchers found that English and twenty-four other languages have enough digital resources for effective LLMs.[170] Another twenty-eight languages, they further observed, "have sufficient text corpora to benefit from the crosslingual zero-shot abilities of LLMs."[171] The researchers called these languages "Rising stars," explaining that "there are a host of opportunities that generative AI has to offer for them that would not have been possible with earlier technologies."[172]

Efforts by various stakeholders, including NGOs, academia, and governments, are dedicated to supporting multilingual capabilities, including the creation of multilingual labeled datasets.[173] These efforts also encompass initiatives such as the Universal Dependencies, an international cooperation aimed at providing an open inventory of categories and guidelines to facilitate consistent labeling across different languages and benchmarks.[174] Universal Dependencies seeks to "encourage more research on multilingual transfer learning," and to

---

[164] *Id.*

[165] Jones et al., *supra* note 157, at 22.

[166] Pires et al., *supra* note 162, at 1.

[167] *Id.* at 1.

[168] OpenAI et al., *GPT-4 Technical Report v6*, ARXIV 1, 7 (Mar. 6, 2023), https://doi.org/10.48550/arXiv.2303.08774 [https://perma.cc/DHT5-52H7].

[169] Choudhury, *supra* note 31, at 1802; *see also* Pratik Joshi et al., *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*, ARXIV (Jan. 17, 2021), https://doi.org/10.48550/arXiv.2004.09095 [https://perma.cc/QS8H-TGTR].

[170] Choudhury, *supra* note 31, at 1802.

[171] *Id.*

[172] *Id.*

[173] Lai et al., *supra* note 15, at § 2.

[174] UNIVERSAL DEPENDENCIES, https://universaldependencies.org/ [https://perma.cc/58RS-VZMS].

https://doi.org/10.59015/wilj.SCTW4297

"maximize language diversity."[175] Another example of such an initiative is XTREME-R benchmark, designed to cover "50 typologically diverse languages spanning 14 language families and 10 challenging, diverse tasks that require reasoning about different levels of syntax, semantics, and common sense."[176]

It is interesting to note that multilingual models not only cater to DMLs but demonstrate better performance in DDLs as well. Carmen Banea et al. stated in this context: "[M]ore languages are better, as they are able to complement each other, and together they provide better classification results. When one language cannot provide sufficient information, another one can come to the rescue."[177] Such benefits can also apply to image understanding, as individuals may differ in visual perception depending on their cultural background and language. Datasets that reflect a variety of languages, therefore, support such informational richness.[178] Indeed, researchers found that models trained on multilingual datasets generated descriptions with higher semantic coverage, on average, compared to models trained on monolingual datasets.[179]

In addition to the efforts around LLMs' multilingualism, various stakeholders focus on developing NLP technologies in local languages. Examples include the Masakhane in Africa, which is an "open-source, continent-wide, distributed, online research effort for machine translation for African languages"[180] and the AI4Bharat in India, a research lab aimed at "developing open-source datasets, tools, models and applications for Indian languages."[181]

---

[175] *Cross-Lingual Transfer Evaluation of Multilingual Encoders*, XTREME, https://sites.research.google/xtreme [https://perma.cc/6M26-TD6S].

[176] *Id.*

[177] Carmen Banea et al., *Multilingual Subjectivity: Are More Languages Better?*, 2010 PROC. 23RD INT'L CONF. ON COMPUTATIONAL LINGUISTICS, 28, 35; *see generally* Anne Aula & Melanie Kellar, 2009 CHI EA '09: CHI '09 EXTENDED ABSTRACTS ON HUM. FACTORS COMPUTING SYS. 3865 (concerning the NLP task of Subjectivity detection).

[178] *Id.*

[179] *Id.* at 1 ("For example, multilingual descriptions have on average 29.9% more objects, 24.5% more relations, and 46.0% more attributes than a set of monolingual captions.").

[180] Iroro Orife et al., *Masakhane — Machine Translation For Africa*, ARXIV 1, 1 (Mar. 13, 2020), https://doi.org/10.48550/arXiv.2003.11529 [https://perma.cc/4CWN-SECQ]; Choudhury, *supra* note 31, at 1803.

[181] *Building AI for India!*, AI4BHĀRAT, https://ai4bharat.iitm.ac.in/ [https://perma.cc/8LAD-LD8S].

https://doi.org/10.59015/wilj.SCTW4297

2. Techno-Social Predicaments

Notwithstanding the advantages offered by LLMs in expanding the digital opportunities available for DML speakers, most DMLs still lag far behind DDLs in various aspects of LLM performance.[182] Moreover, LLMs might be biased against DML speakers in different fashions, as will be discussed below.[183]

Surangika Ranathunga and Nisansa de Silva observed that the majority of languages in the world—numbering in the thousands as aforementioned—"have received limited to no attention" from NLP technologies.[184] Siavoshi explained that most NLP processes only focus on English and a few additional languages,[185] leaving LLMs' capacities in other languages, including minority and indigenous languages, lacking.[186]

Such inequalities apply to both multilingual and language-specific models.[187] Concerning the former, Doddapaneni et al. stated that despite their potential to mitigate the performance gap between high- and low-resource languages through zero-shot learning, "in practice, the benefits of such models are still skewed towards high-resource languages."[188] Microsoft India research—noted above as indicating positive NLP developments for a confined group of languages—also warned that for "the remaining 99% of the world's languages—spoken by approximately 1.5 billion people—LLMs have little to offer."[189]

Back tracing the roots of these limitations, and meaningfully exploring their implications are not easy tasks. Ranathunga and de Silva noted that the reasons behind existing linguistic disparities "are seldom

---

[182] *E.g.* Lai et al., *supra* note 15; *see generally* Philipp Ennen et al., *Extending the Pre-Training of BLOOM for Improved Support of Traditional Chinese: Models, Methods, and Results*, ARXIV (Jun. 23, 2023), https://doi.org/10.48550/arXiv.2303.04715 [https://perma.cc/NWD6-AS9D].

[183] *See infra* Part II.B.2.

[184] Ranathunga & de Silva, *supra* note 13 at 1; Peter Baumann & Janet Pierrehumbert, *Using Resource-Rich Languages to Improve Morphological Analysis of Under-Resourced Languages*, *in* PROC. NINTH INT'L CONF. ON LANGUAGE RES. & EVALUATION 3355 (2014) (declaring that the number of languages for which NLP systems are available "is small compared to the nearly 7000 languages spoken on the planet").

[185] Siavoshi, *supra* note 35.

[186] Navigli et al., *supra* note 75, at 10.

[187] See *infra* this Part.

[188] Doddapaneni et al., *supra* note 32, at 12402.

[189] Choudhury, *supra* note 31, at 1802; *see also* Orife et al., *supra* note 181 ("[F]ew efforts to fund NLP or translation for African languages exist, despite the potential impact. This lack of focus has had a ripple effect").

https://doi.org/10.59015/wilj.SCTW4297

discussed within the NLP community."[190] In addition, critically experimenting with deep learning, the technology on which LLMs are based, often requires large-scale data. As such, insights regarding training data have become "concentrated within a few organizations, many of which do not openly share their findings and methodologies."[191]

Nonetheless, I aim to outline several factors that drive this troubling asymmetry, while exploring both technological and sociocultural considerations. I will start by discussing LLMs' inequalities around training data and training processes and proceed with the linguistic gaps in LLMs' design and evaluation choices and limitations.

a. Training Data and Training Processes

One significant cause for the linguistic gaps between DMLs and DDLs is that DML data is often absent from LLM training processes. To a great extent, this builds on the under-representation of DMLs in LLMs' unlabeled and labeled datasets.[192] Though information about the data used for training LLMs is not often disclosed by the private companies that developed them,[193] several points become clear and should be considered.

First, the vast majority of the world's languages do not have enough available training data to power a language-specific LLM.[194]

Second, most of the world's languages are absent from multilingual pretraining settings of LLMs.[195] Often, these models only encompass a few tens of languages[196] (or around one hundred languages in several other cases).[197] Microsoft India research, noted above, calls the excluded languages "the left-behinds," explaining that they "have been and are still ignored in the aspect of language technologies. . .Unsupervised pre-training methods only make the 'poor

---

[190] Ranathunga & de Silva, *supra* note 13, at 1.

[191] Alon Albalak, *A Survey on Data Selection for Language Models,* arXiv (Aug. 2, 2024), https://doi.org/10.48550/arXiv.2402.16827 [https://perma.cc/8J4T-ZZA6].

[192] Wenhao Zhu et al., *Extrapolating Large Language Models to Non-English by Aligning Languages*, arXiv (Oct. 9, 2023), https://arxiv.org/pdf/2308.04948 [https://perma.cc/ECP8-MT74]; *see also* European Parliamentary Research Service, Language Equality in the Digital Age: Towards a Human Language Project (2017).

[193] Navigli et al., *supra* note 75, at 10.

[194] *E.g.* Choudhury, *supra* note 31, at 1802.

[195] Doddapaneni et al., *supra* note 32, at 12402; *see also* Ebrahimi et al., *supra* note 152, at 6279 (stating, with relation to low-resource languages, that they are "most likely to be unseen to commonly used pretrained models. . .").

[196] Yuan et al., *supra* note 161, at 2.

[197] Navigli et al., *supra* note 75, at 10.

https://doi.org/10.59015/wilj.SCTW4297

poorer,' since there is virtually no unlabeled data to use."[198] This state of affairs builds on the important role that internet-based resources—which primarily consist of DDLs and particularly English—play in the training of LLMs. One example for such resources is Common Crawl, the standard dataset for pretraining data.[199]A further troubling picture emerges in the context of the fine-tuning stages, which require labeled data, since DML linguistic labeling is often partial or absent from relevant data repositories.[200]

Third, for the limited number of DMLs that did manage to find their way to the pretraining data of multilingual language models, these languages only constitute a small fraction of the entire dataset used in the pretraining stage.[201] Doddapaneni et. al. explained that "due to this disparity, low-resource languages get a very poor share of the model's capacity and vocabulary, and thus the performance on these languages is poor."[202] Navigli et al. noted that "it is not surprising that a multilingual system trained on an unbalanced distribution of languages will perform better in those languages for which the training data was richer in quantity and quality."[203]

Indeed, most of the data used for training multilingual LLMs is often in English. Llama, for instance, includes only 4.5 percent multilingual data, whereas the rest is English data. Llama2 demonstrates an advancement in this regard, with about 11 percent multilingual data.[204] ChatGPT-3's unlabeled pretraining data spans 119 languages, but they only account for about 7 percent of the tokens/words included. The remaining 93 percent are English tokens/words.[205] Other LLMs—such as

---

[198] Joshi et al., *supra* note 169, at § 2.3.

[199] Agasthya Gangavarapu, *LLMs: A Promising New Tool for Improving Healthcare in Low-Resource Nations*, *in* 2023 IEEE GLOB. HUMANITARIAN TECH. CONF. 252, 253 (2023) (warning that this "neglects other languages like Hindi, the third most spoken language globally, which accounts for a mere 0.15% of the corpus, leading to significant underrepresentation"); *see Statistics of Common Crawl Monthly Archives*, GITHUB, https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html [https://perma.cc/DN3J-SB34] (listing the languages included in the Common Crawl dataset).

[200] Joshi et al., *supra* note 169, at § 2.4; s*ee also* Baumann & Pierrehumbert, *supra* note 184, at 3355 (stating that most languages are "lacking not only practical NLP systems, but even the large labeled corpora typically used to develop such systems").

[201] Doddapaneni et al., *supra* note 32, at 12402–03.

[202] *Id.* at 12403.

[203] Navigli et al., *supra* note 75, at 6.

[204] Yuan et al., *supra* note 162, at 2.

[205] Kabir Ahuja et al., *MEGA: Multilingual Evaluation of Generative AI*, ARXIV (Oct. 22, 2023), https://doi.org/10.48550/arXiv.2303.12528 [https://perma.cc/4Q5J-HLPW]; *Openai/GPT-*

PaLM, whose training data includes about 22 percent non-English text—offer a better representation of languages.[206] Even so, these rates still reflect a clear disadvantage for DMLs vis-à-vis DDLs.

Expanding the multilingual pre-trained data is desirable, inter alia, since, in some cases, even a small increase in the diversity of data considerably improves the performance of LLMs in various languages.[207] Other attempts to improve LLMs' capacities across languages involve including DMLs in the fine-tuning stages. This is not always a simple task as it requires labeled data, a resource which is often scarce for these languages.[208]

These concerns are summarized in the words of Wenhao Zhu et al.:

> The language ability of LLMs is often imbalanced across languages. . .because both the pre-training corpus. . .and the instruction-tuning data. . .are English-dominated. As a result, LLMs usually perform poorly on non-English languages, especially on languages that are dissimilar to English.[209]

Though zero-shot learning enables advancements for DMLs, since it does not involve fine-tuning, and therefore, does not require labeled data,[210] it should be regarded with a grain of salt.[211]

Niklas Muennighoff et al. noted: "zero-shot performance tends to be significantly lower than finetuned performance. Thus, task-specific or language-specific transfer learning via finetuning remains the predominant practice. . .This is particularly challenging for low-resource languages or tasks with limited data available."[212] Zero-shot learning also raises particular concerns regarding models' ability to perform high-level NLP tasks that require reasoning.[213]

3Public Archive, GITHUB.COM, https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv [https://perma.cc/96SM-VVC6].

[206] Aakanksha Chowdhery et al., *PaLM: Scaling Language Modeling with Pathways*, ARXIV 1, 32 (Oct. 5, 2022), https://doi.org/10.48550/arXiv.2204.02311 [https://perma.cc/D2DH-8DK5].

[207] Uri Shaham et al., *Multilingual Instruction Tuning With Just a Pinch of Multilinguality*, ARXIV (Jan. 3, 2024), https://doi.org/10.48550/arXiv.2401.01854 [https://perma.cc/DY8Y-9TEX].

[208] *See supra* Part II.B.

[209] Zhu et al., *supra* note 193; *see also* Baumann & Pierrehumbert, *supra* note 184, at 3355.

[210] *See supra* Parts II.A. and II.B.1.

[211] Niklas Muennighoff et al., *Crosslingual Generalization through Multitask Finetuning*, ARXIV (May 29, 2023), https://doi.org/10.48550/arXiv.2211.01786 [https://perma.cc/8P69-E2C5]; *see also* Lauscher et al., *supra* note 139.

[212] Muennighoff et al., *supra* note 211, at 1.

[213] Ebrahimi et al., *supra* note 151, at 6279, 6286–87; *see supra* Part II.A. (addressing the difference between high and low-level NLP tasks).

https://doi.org/10.59015/wilj.SCTW4297

Moreover, zero-shot learning's limitations for DMLs also stem from linguistic cross-family implications, meaning that differences between language families pose challenges to models' performance across diverse linguistic systems.[214] Some LLMs' transfer learning is conditioned upon similarity in typology (word order) between the original and target languages.[215] In addition, some LLMs' performance improves when there is a high lexical overlap between the paired languages—that is, when they share a significant amount of similar vocabulary.[216]

Thus, in (the likely) cases where most of the training data is in English, for instance, transfer learning for different and remote language families may lead to disfavored performance. Another factor that reduces zero-shot learning's performance is the scarcity of DML data in the pre-trained corpora, a common situation in multilingual LLM training, as discussed above.[217]

Fourth, the composition of the training data, since dominated by DDLs, will often not carry DML speakers' narratives, lived experiences, values, histories, and needs.[218] I have addressed the importance of original content in DMLs earlier when discussing digitalization inequalities in general.[219] However, I find it worthwhile to further explore this concern directly in the context of LLMs and NLP. Mehrnaz Siavoshi emphasized in this regard:

> As a relatively agnostic system is trained on English, it learns the norms and systems of a specific language and all the cultural implications that come with that limitation. This single-sided approach will only continue to become more apparent as NLP is applied to more intelligent processes that have an international audience.[220]

It was also noted that "an in-built language preference in the systems that drive the internet inherently incorporates the societal norms of the driving languages."[221] In the same vein, Dodge et al. pointed out that

---

[214] Doddapaneni et al., *supra* note 32, at 12403 ("The ability of multilingual models to do zero-shot transfer is often limited to typological cousins inside language families. . .").

[215] Pires et al., *supra* note 162, at 1-2; *see also* Lauscher et al., *supra* note 139.

[216] Pires et al., *supra* note 162, at 1.

[217] Lauscher et al., *supra* note 139, at 2. More specifically, it was found that Zero-shot learning reduces the models' performance in low-level NLP tasks if there is no structural similarity between the source and target languages, and that low representation in the pretraining corpora adversely affects high-level language understanding tasks. *See supra* note 119 and accompanying text.

[218] See *infra* note 220 and accompanying text.

[219] *See supra* Part I.B.

[220] Siavoshi, *supra* note 35.

[221] *Id.*

https://doi.org/10.59015/wilj.SCTW4297

languages, including metaphors, idioms, and figurative expressions, reflect their speakers' cultures.[222] Using a "skewed distribution" of languages in computational models, they explained, results in gaps in cultural representation.[223] Furthermore, they noted that since different populations discuss different topics, creating more linguistically inclusive models can help reduce biases toward the values associated with better-represented languages.[224]

This challenge is further underscored as a significant part of website domains whose content is used for training are US domains and other English-speaking countries' domains, including Canadian, Australian, and UK domains.[225] Some leading websites, in this regard, include American news outlets, such as The New York Times and The Washington Post. The UK Guardian also occupies a leading place in the training data.[226] Other relevant repositories are Reddit (where much of the content is created by Americans), and English Wikipedia.[227] All these sources may represent only a confined set of values and agendas and thus exclude those of many DML communities.

Finally, another concern that must be noted is that text generated by AI will probably become increasingly present in the future training data of LLMs.[228] Indeed, it was noted that "[a]s the use of models which can generate natural language text proliferates, web-crawled data will increasingly contain data that was not written by humans."[229]

This phenomenon may exacerbate the asymmetries between DMLs and DDLs, reinforcing a linguistic exclusion cycle.[230] Given the

---

[222] *See* Jesse Dodge et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus,* PROCEEDINGS OF THE 2021 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1286, 1293-94 (2021).

[223] *Id.* at 1292.

[224] *Id.* at 1291.

[225] *Id.* at 1288. The leading domain list also includes, albeit to a lesser degree, domains where English is not the dominant language. Most of these are, however, administered in Europe (including the European Union domain).

[226] *Id.* Other news outlets, such as Al Jazira, are also included in the list but ranked lower.

[227] Rettberg, *supra* note 193.

[228] Charley Johnson, *AI Companies Are Running Out of High-Quality Data. Here's What Happens Next.*, UNTANGLED WITH CHARLEY JOHNSON (Apr. 14, 2024), https://untangled.substack.com/p/ai-companies-are-running-out-of-high [https://perma.cc/WR55-MRK5].

[229] Dodge, *supra* note 222, at 1289; *see also* Sina Alemohammad et al., *Self-Consuming Generative Models Go MAD*, ARXIV (Jul. 4, 2023), https://doi.org/10.48550/arXiv.2307.01850 [https://perma.cc/S8BL-B623].

[230] *Content*, MEDIUM (Jul. 14, 2023), https://towardsdatascience.com/ai-entropy-the-vicious-circle-of-ai-generated-content-8aad91a19d4f [https://perma.cc/7H88-KJPJ].

https://doi.org/10.59015/wilj.SCTW4297

limited amount of organic training data[231] and the vast amount of AI-generated data—such as GPT-3's daily output of approximately 4.5 billion words,[232] which is expected to increase further[233]—this concern merits serious consideration. In addition, some AI-generated data used for future training might be content translated to DMLs, rather than originally created in such languages.[234] Translation, as noted earlier, may transmit various cultural biases existing in the source language, and not represent the topics and concerns most relevant to DML communities.[235] This might be, thus, another source for cementing cultural biases in LLMs' operations.

All these considerations reflect the importance of linguistically diversifying the pretraining and fine-tuning processes. Non-inclusive training data may lead to reduced performance in DMLs across different NLP tasks,[236] and to the generation of excluding and discriminating content, as explored above.[237] Since LLMs are incorporated in a wide range of applications, products, and services, such insufficient and biased capabilities may carry broad ramifications across different facets of life.[238]

b. Design and Evaluation Choices and Constraints

Additional sources of linguistic limitations and biases against DMLs include design and evaluation choices and constraints.[239] Researchers discussed, for instance, how current tokenization processes favor Latin and Cyrillic European languages over other linguistic families. Tokenization is the process of segmenting text into smaller components called tokens, before it is fed to language models. In DDLs such as English and Spanish, the tokenizer recognizes most words as a single token, while in some DMLs such as Tamil or Thai, it may break the words down into smaller components. Since the "context window"—the number of tokens that LLMs can process in a given time—is limited, this tokenization

---

[231] Johnson, *supra* note 228.

[232] *PT-3 Powers the Next Generation of Apps*, OPENAI (Mar. 25, 2021), https://openai.com/blog/gpt-3-apps [https://perma.cc/Y6PA-UFGP].

[233] *Id.*

[234] *See* Josh Emanuel, *The Looming Crisis of Web-Scraped and Machine-Translated Data in AI-Language Training*, APPEN, (April 4, 2024), https://www.appen.com/blog/web-scraped-and-machine-translated-data-in-ai-language-training [https://perma.cc/KJ9Y-YYYL].

[235] *See supra* Part I.A.II.

[236] *See supra* Part II; *see also* Chowdhery et al., *supra* note 206, at 32; Siavoshi, *supra* note 35 (regarding the reduced performance).

[237] *See supra* Part II.

[238] *See supra* Part II.B.1.

[239] Choudhury, *supra* note 31, at 1802–03.

https://doi.org/10.59015/wilj.SCTW4297

architecture entails that LLMs may process less DML data in comparison to DDL data.[240] Among other consequences, this may reduce the length of DML documents that LLMs can handle. One solution to this concern may involve the use of language-specific tokenizers.[241]

Another design-driven difficulty concerns the filtering of data used for training LLMs. Concerns have been raised that these filtering practices might disproportionately remove nonstandard English, such as so-called African American English and Hispanic-aligned English, more than other English versions and dialects.[242] Research also indicates that filtering may adversely censor content written about and by minority groups.[243] One of the reasons for this is that these linguistic variations may be identified as grammatically incorrect and low-quality data.[244] This may have additional biased results. It was explained, for instance, that "a direct consequence of removing such text from datasets used to train language models is that the models will perform poorly when applied to text from and about people with minority identities, effectively excluding them from the benefits of technology like machine translation or search."[245] A more calibrated and cautious approach to data filtering could reduce the severity of this problem.[246]

Another design choice concerns the evaluation sets (validation sets and test sets) selected by developers to assess the generalization and transfer learning capabilities of language models. Viet Dac Lai et al. noted that "ChatGPT has been mainly evaluated over English data. The community is lacking a comprehensive, public, and independent evaluation of ChatGPT over various non-English languages for diverse NLP tasks to provide proper perspectives for future research and

---

[240] *Id.* at 1802. For efforts to extend the context window, see, for instance, Yiran Ding et al., *LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens*, ARXIV, HTTPS://ARXIV.ORG/ABS/2402.13753 [HTTPS://PERMA.CC/V5YP-KU47].

[241] *Id.*

[242] Dodge, *supra* note 222, at 1292. Such filtering practices may also disproportionately remove content relating to LGBT+, among other vulnerable groups. *See id*; BigScience Workshop et al., *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*, ARXIV 1, 9 (Jun. 27, 2023), https://arxiv.org/abs/2211.05100 [https://perma.cc/VZ86-2XUH].

[243] Albalak et al., *supra* note 191, at 44.

[244] Barbara Plank, *What to Do About Non-Standard (or Non-Canonical) Language in NLP,* ARXIV 1, 1 (Aug. 28, 2016), https://arxiv.org/pdf/1608.07836 [https://perma.cc/XKL8-HJ3C].

[245] Dodge, *supra* note 222, at 1293.

[246] *See* Steinþór Steingrímsson et al., FILTERING MATTERS: EXPERIMENTS IN FILTERING TRAINING SETS FOR MACHINE TRANSLATION, PROCEEDINGS OF THE 24TH NORDIC CONFERENCE ON COMPUTATIONAL LINGUISTICS 588–600 (NoDaLiDa, 2023).

https://doi.org/10.59015/wilj.SCTW4297

applications."[247] Choudhury explained that "[d]evelopers often seek to improve the average performance of models by using evaluation sets that are skewed towards high-resourced languages," suggesting that they should instead "focus on minimizing the performance gap between the lowest- and highest-performing languages."[248]

A final group of factors that powers the discrepancies between DMLs and DDLs concerns benchmarks. Currently, there is a lack of benchmarks in languages other than English, and especially in DMLs. Despite the advancements reflected in benchmarks such as the aforementioned XTREME-R,[249] they still do not cater to the vast majority of languages in the world, particularly the vulnerable ones. Sumanth Doddapaneni et al. stated: "15 of the 22 constitutionally recognized Indic languages have no representation in XTREME-R for any task."[250] Researchers[251] and private AI companies[252] sometimes translate benchmarks to English or other DDLs,[253] but this practice may bake in translation mistakes as well as other biases, thereby further deepening the gap between DMLs and DDLs.[254]

III. LLMS, DIGITAL SIDELINING, AND PARTICIPATION

A. "PARITY OF PARTICIPATION"

As previously discussed, linguistic sidelining has long-lasting and powerful roots.[255] Digitalization—including LLMs—offer invaluable advancements to DMLs. Nonetheless, they also introduce significant concerns, including biases and low performance in these languages.

These challenges largely stem from the training data and training processes, as well as the design choices and constraints relevant to these

---

[247] Lai et al., *supra* note 15, at 2.
[248] Choudhury, *supra* note 31, at 1802–1803.
[249] *See supra* Part II.B.1.
[250] Doddapaneni et al., *supra* note 32, at 12403.
[251] *See* Ahuja et al., supra note 205.
[252] OpenAI et al., *supra* note 168, at 7.
[253] *Id.*
[254] *See supra* Part I.A.
[255] *See supra* Part I.

https://doi.org/10.59015/wilj.SCTW4297

technologies.[256] Digitalization, and specifically LLMs, may, therefore, reinforce and deepen existing inequalities between DDLs and DMLs.[257]

One central challenge this linguistic sidelining creates is preventing DML speakers from meaningfully participating in our current, intensely digital society. As non-participants, DML speakers may be ill-equipped, passive, and even transparent actors in spheres designed to serve DDLs first and foremost.[258] Indeed, Tommaso M. Milani et al. noted that language is "a gatekeeper for participation" and that "language choice is one of the structural components that enables or hinders participation."[259]

To better understand the nature, scope, and implications of the participation deprivation experienced by DML speakers in digital avenues, I build on Nancy Fraser's influential writing. Though originally focusing on offline linguistic challenges, I find Fraser's insightful theoretical framework also beneficial in our digital context.

Fraser perceives "Parity of Participation" as reflecting the most general meaning of justice.[260] Parity of Participation, she clarified, requires "social arrangements that permit all to participate as peers in social life."[261] She further noted that such equality can only be achieved if

> all the relevant subjects have no entrenched social obstacles that in a structural way prevent them from participation in terms of parity or equality—whether this is participation in formal and informal political and public spheres, institutions, life, in civil society, in the life of associations, in family life, in labour markets, in fact in any and all of the major institutional arenas that are important in society.[262]

According to Fraser, participation builds on three layers: distribution, recognition, and representation.[263] Maldistribution lies in economic structures that deny people the resources required to interact with others as peers.[264] Distributive justice seeks equality in the allocation of such resources, which may include "rights,. . .primary goods, opportunities, real freedoms, and capabilities."[265] Misrecognition relates

---

[256] *See supra* Parts I and II.

[257] *See supra* Parts I and II.B.

[258] *See supra* Parts I and II.B.2.

[259] Tommaso Milani et al., *Participation on Whose Terms? Applied Linguistics, Politics and Social Justice*, 2023 AFINLAN VUOSIKIRJA 287, 287.

[260] FRASER, *supra* note 37, at 16.

[261] *Id.*

[262] Chhachhi*, supra* note 39, at 303.

[263] FRASER, *supra* note 37, at 16–18.

[264] *Id.*

[265] *Id.* at 32.

https://doi.org/10.59015/wilj.SCTW4297

to institutionalized hierarchies of cultural values and status.[266] Conversely, recognition aspires to "a world where assimilation to majority or dominant cultural norms is no longer the price of equal respect."[267] "What requires recognition," Fraser clarified, "is not group-specific identity but rather the status of group members as full partners in social interaction."[268]

Misrepresentation, which concerns the third layer in Fraser's framework, pertains to the exclusion of people and groups from decision-making processes.[269] Misrepresentation includes both ordinary political representation and meta-political representation, which Fraser terms "misframing." The former component of representation concerns the procedures of decision-making while misframing addresses "the boundary setting aspect" of the political landscape, that is, the "inclusion in, or exclusion from, the community of those entitled to make justice claims on one another."[270] In other words, the ordinary political aspect of representation concerns the rules according to which decisions are made, while framing dictates who gets to take part in the decision-making.[271] Instead of the residency in a territorial state, Fraser embraced the "all-affected principle" as the rule determining the circle of participants relevant to different globalizing concerns. According to this principle, "all those affected by a given social structure of institution have moral standing as subjects of justice in relation to it." She noted that "what turns a collection of people into fellow subjects of justice is not geographical proximity, but their co-imbrication in a common structural or institutional framework, which sets the ground rules that govern their social interaction,

---

[266] *Id.* at 16–18.

[267] Nancy Fraser, *Recognition without Ethics?*, 18 THEORY, CULTURE & SOC'Y 86, 86 (2001).

[268] *Id.* at 89.

[269] FRASER, *supra* note 37, at 16–18.

[270] *Id.* at 17; *see also id.* at 18–21, 19 ("When questions of justice are framed in a way that wrongly excludes some from consideration, the consequence is a special kind of meta-justice, in which one is denied the chance to press first-order justice claims in a given political community.").

[271] Misrepresentation, Fraser explains, "occurs when political boundaries and/or decision rules function wrongly to deny some people the possibility of participating on a par with others in social interactions - including, but not only, in political arenas." *Id.* at 18. Demands for reframing, according to Fraser, can be affirmative, in that they accept the validity of the basic unit at hand–specifically in her writing: the territorial state–and only demand that this unit's boundaries will be drawn differently, thus influencing the group of people who participate in decision making. Calls for reframing can also be transformative, Fraser explains. Proponents of this approach do not entirely oppose the state-territoriality. Instead, "they contend that its grammar is out of sync with the structural causes of many injustices in a globalizing world, which are not territorial in nature." One example discussed by Fraser in this regard is the "information networks of global media and cybertechnology, which determine who is included in the circuits of communicative power and who is not." *Id.*

https://doi.org/10.59015/wilj.SCTW4297

thereby shaping respective life possibilities in patterns of advantage and disadvantage."[272]

Together, distribution, recognition, and representation constitute three central contexts necessary for meaningful, sustainable, and robust participation across various facets of life. These pillars influence key aspects of the economic, cultural, and governance structures that systematically—though sometimes seamlessly—deny DML speakers of such equal opportunities for participation. Regrettably, these three pillars are often challenged in digital settings.

B. MIS-PARTICIPATION IN THE DIGITAL LINGUISTIC CONTEXT

I will now examine the application of each of Fraser's theoretical layers to DMLs, with a primary focus on the distribution level, due to its close connection to the preceding techno-social discussion.

1. Maldistribution

Existing biased economic structures deny DMLs access to the wealth of resources available to DDL speakers. These structures encompass financial disincentives to diversify the digital linguistic ecosystem, invest in inclusive training sets and training processes, and reassess the design and evaluation of LLMs so that they will also adequately serve DMLs rather than only DDLs. Navigli et. al. addressed the vicious circle these economic structures generate, in which DDL domination and DML exclusion are further entrenched:

> It is undeniable that most of the work in NLP revolves around high-resource languages. The reason is obvious. For a high-resource language L, collecting data and hiring linguists and annotators is easier; this situation has enabled a vicious cycle in which it is simpler to develop an NLP system for L and identify new challenges to work on within the scope of L, leading to the creation of more data for L and, in turn, to the development of better systems for L.[273]

This unfair allocation of resources further blocks DML individuals and communities from enjoying the benefits, human rights, and opportunities provided by and dependent upon digitalization.

---

[272] *Id.* at 24.

[273] Navigli et al., *supra* note 75, at 6.

https://doi.org/10.59015/wilj.SCTW4297

At the individual level, DML speakers may face barriers in accessing informational and communication venues, critically formulating ideas and opinions, and engaging in local and global discourses.[274] This could hamper their prospects of societal and financial mobilization.[275] Linguistic maldistribution may also limit DML speakers' ability to enjoy a broad range of human rights, including the right to education,[276] health,[277] culture,[278] dignity, well-being,[279] and identity.[280] Indeed, language and identity, Ruth Wodak has noted, "have a dialectic relationship. Languages manifest 'who we are,' and we define reality partly through our language and linguistic behavior."[281] DML speakers may also not have the resources to choose or challenge religious, political, and societal assumptions and phenomena, thus limiting their freedom of thought, religion, and even their right to vote.[282]

Unequal distribution, in our case, not only distances DMLs from valuable opportunities but may also expose them to unjustified sanctions and erroneous decisions, whether made by governments or digital platforms. Content moderation processes are one example of an area where such sanctions may apply.[283] NLP and LLM-based moderation processes might miss out on linguistic nuances or political and social sensitivities

---

[274] JONES ET AL., *supra* note 157, at 22.

[275] *See generally* Irene de Zarzà et al., *Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations*, 5 AI 91 (2024).

[276] *See, e.g.*, Wensheng Gan et al., *Large Language Models in Education: Vision and Opportunities*, 2023 IEEE INT'L CONF. ON BIG DATA (BIGDATA) 4776 (2023), https://arxiv.org/pdf/2311.13160 [https://perma.cc/9SDG-WU3G]; Qingyao Li et al., *Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges*, Arxiv.org (2024), https://arxiv.org/pdf/2401.08664 [https://perma.cc/ZR5K-HE9Q].

[277] Jorge A. Rodriguez et al., *Leveraging Large Language Models to Foster Equity in Healthcare*, 31 J. AM. MED. INFORMATICS ASS'N 2147, 2147 (2024).

[278] *See, e.g.*, Trichopoulos, *supra* note 134.

[279] Ranathunga & de Silva, *supra* note 13, at 1–2.

[280] Ruth Wodak, *Language, Power and Identity*, 45 LANGUAGE TEACHING 215, 216 (2012); *see also* Dafna Dror-Shpoliansky & Yuval Shany, *It's the End of the (Offline) World as We Know It: From Human Rights to Digital Human Rights – A Proposed Typology*, 32 EUR. J. INT'L. L. 1249 (2021) (discussing human rights in the digital age).

[281] Wodak, *supra* note 280, at 216.

[282] *See* Kebene Wodajo, Realising the Societal Dimensions of the Right to Freedom of Thought in the Digital Age Through Strategic Litigation, *in* CAMBRIDGE HANDBOOK RT. TO FREEDOM OF THOUGHT 363, 363–64 (forthcoming 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4754468 [https://perma.cc/2S3P-HQPW] (discussing digitalization in general).

[283] *See* Troy Wolverton, *AI is Great at Recognizing Nipples, Mark Zuckerberg Says*, BUS. INSIDER (Apr. 25, 2018, 5:47 PM), https://www.businessinsider.com/ai-can-identify-nipples-but-not-hate-speech-mark-zuckerberg-says-2018-4 [https://perma.cc/779K-VTJL].

https://doi.org/10.59015/wilj.SCTW4297

relevant to DMLs. These processes may mistake criticism, reclaiming, or sarcasm for offensive content, leading to this content's removal, and even to the deplatforming of its authors.[284] Sanctions might also be applied by law enforcement and other governmental bodies and contractors, which increasingly incorporate AI in different areas.[285]

At the communal level, the underlying economic mechanisms in the LLM context may prevent entire DML-speaking communities from fully enjoying their culture, celebrating their language, and uniting around it.[286] Laurent Besacier et al. explained that "both, individual and community memories, ideas, major events, practices, and lessons learned are all preserved and transmitted through language,"[287] and David Crystal has noted that local languages enhance communal cohesion and stimulate a sense of pride and confidence within communities.[288]

Indeed, "the intimate relationship between language and culture is widely recognized."[289] This approach is further underscored in the words of a Nigerian formal representative, explaining that "once the language dies, the culture dies."[290] Along similar lines, because languages are key expressions of minority groups' identities, the disappearance of a language is likely to be accompanied by the disappearance of the associated minority group.[291]

---

[284] *The Consequences of Meta's Multilingual Content Moderation Strategies,* DIGITAL WATCH OBSERVATORY (Sep. 1, 2023), https://dig.watch/updates/the-consequences-of-metas-multilingual-content-moderation-strategies [https://perma.cc/3R5K-ZD35]; *see also* Wolverton, *supra* note 283.

[285] *See* Warner, *supra* note 41; *see also* Biron, *supra* note 41. It was further stated: "The machines themselves are not operating with even a fraction of the quality they need to be able to do case work that's acceptable for someone in a high-stakes situation." *Id.*; Paria Sarzaeim et al., *A Systematic Review of Using Machine Learning and Natural Language Processing in Smart Policing*, 12 COMPUTERS 255 § 4.1.3 (2023).

[286] Burn et al., *supra* note 48, at 377.

[287] Besacier et al., *supra* note 27, at 87.

[288] DAVID CRYSTAL, LANGUAGE DEATH 31 (2002).

[289] Ruth Rubio Marin, *Language Rights: Exploring the competing rationales*, *in* LANGUAGE RTS. & POL. THEORY 52 (Will Kymlicka and Alan Patten, eds., 2003).

[290] Rita Izsák (Special Rapporteur), Hum. Rts. Council, Report of the Special Rapporteur on Minority Issues, at 17, U.N. Doc. A/HRC/28/64/Add.2 (Jan. 5, 2015).

[291] SILVIA QUATRINI, A RIGHTS-BASED FRAMEWORK FOR MINORITY AND INDIGENOUS LANGUAGES IN AFRICA: FROM ENDANGERMENT TO REVITALIZATION (2019). It is important to note that when resources, freedoms, and opportunities are systematically prevented from DML-speaking individuals and communities, harm is also inflicted upon society at large, endangering core democratic values, such as equality and diversity. Encouraging an inclusive society reduces fragmentation and stigmatization among different sectors of society. It may dilute our cultural richness and the wealth of the available perspectives, beliefs, histories, and values. *See, e.g.*, Besacier et al., *supra* note 27, at 87. Interestingly, David Crystal has found similarities between language diversity and biological diversity. He explained that multilingualism is a component of

https://doi.org/10.59015/wilj.SCTW4297

2. Misrecognition

In addition to distribution, recognition—the second layer of Fraser's "Parity of Participation" framework—is also insufficiently available for DML speakers.

While the discussion on distribution illuminates, inter alia, how DML are sidelined, misrecognition concerns the cultural asymmetries on which such maldistribution rests. Indeed, the technological challenges only cover a part of DML speakers' disfavoring. Seeta Peña and Jędrzej Niklas warn that a "techno-centric" approach that leans too heavily on technological adjustments as facilitators of justice, misses out on the broader systemic nature and manifestations of discrimination.[292] The sociocultural aspects of linguistic discrimination are explored in a rich body of literature, such as Tove Kuttab-Kangas's notion of "linguicism," discussed above.[293] The institutionalized cultural devaluation of DMLs is highlighted against the backdrop of the enduring linguistic oppression processes that predate digitalization.[294] It is further embodied in the poor performance of LLMs in DMLs, and the absence of many DMLs from the entire LLM development pipeline, which includes DDL-centered training processes and datasets, evaluation sets and benchmarks, and discriminative data filtering approaches.[295]

LLM-driven translation of DML content to DDLs, despite its benefits, cannot adequately advance recognition, as conceptualized in Fraser's framework. This is because such translation necessitates assimilation to DDLs as the cost of accessing digitalization benefits.[296] Moreover, translated data and zero-shot transfer learning, which mainly rely on DDL data, do not necessarily embody the topics, viewpoints, and values of DML communities, thereby reinforcing their unequal cultural status.[297] Indeed, Choudhury stated that "even when LLMs produce fluent text in several non-European languages, the outputs are often culturally,

---

society's sustainability and that through language extinction, "a serious loss of inherited knowledge" occurs. *See* CRYSTAL, *supra* note 288, at 34.

[292] *See generally* Seeta P. Gangadharan & Jędrzej Niklas, *Decentering Technology in Discourse on Discrimination*, 22 INFO., COMMC'N. & SOC'Y. 882 (2019).

[293] *Supra* Part I.B.

[294] *Supra* Part I.B.

[295] *See supra* Part II.B.2.

[296] Fraser, *supra* note 267, at 21.; *see also supra* Part III.A.

[297] Fraser, *supra* note 267, at 21.; *see also supra* Part III.A.

https://doi.org/10.59015/wilj.SCTW4297

morally and epistemologically misaligned owing to the predominantly Western and Anglocentric representation space."[298]

This systematic cultural marginalization sustains DML speakers' place as insignificant actors in digitalization.[299] DMLs are so inherently sidelined in digital domains that their exclusion might seem, at times, a given by-product of modernization. In this inherently unequal environment, DML speakers, and society at large, might be nudged or expected to be satisfied with any digital advancement offered in connection to DMLs.[300]

This state of affairs may also carry adverse future implications for DML speakers' recognition, as NLP and LLM-generated content in various areas flows back into our informational settings and is further reused as training data.[301]

Finally, it should be noted that many dominant AI platforms are not demonstrating sufficient accountability concerning the risks they pose, including their potential adverse impact on DMLs.[302] For instance, unlike many social media services, AI platforms do not submit "transparency reports," which are periodic publications addressing various aspects of their performance.[303] This opacity creates obstacles to identifying linguistic gaps, raising public awareness about the pertinent cultural asymmetries, and mitigating them.

3. Misrepresentation

Along with maldistribution and misrecognition, misrepresentation is also evident in digital linguistic contexts. This is reflected, inter alia, in the absence of DML speakers from decision-making processes that shape digitalization, despite these decisions having far-reaching implications for them.[304]

---

[298] Choudhury, *supra* note 31, at 1082.

[299] *See supra* Parts I, II.B.2.

[300] *Id.*

[301] *See supra* Part II.B.2.

[302] Rishi Bommasani et al., *The Foundation Model Transparency Index,* Arxiv.org, 1, 29–37 (2023), https://doi.org/10.48550/arXiv.2310.12941 [https://perma.cc/MW8E-T66N]; *see also* Rishi Bommasani et al., *Foundation Model Transparency Reports,* Arxiv.org 1, 1 (2024) https://doi.org/10.48550/arXiv.2402.16268 [https://perma.cc/K88Q-WJZH].

[303] *See* Bommasani et al., *The Foundation Model Transparency Index, supra* note 302 (calling for AI platforms to embrace transparency reports also in the language models' context).

[304] *See infra* Part III.B.3.

https://doi.org/10.59015/wilj.SCTW4297

Consider the AI governance settings in the US, one of the most influential countries in the AI industry, including LLM technologies.[305] Google (Gemini), Microsoft and OpenAI (ChatGPT), Anthropic (Claude), and Meta (Llama), to name a few of the globally dominant actors, are all American companies, with their impact expanding far beyond the US.[306]

The White House Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, signed in October 2023 (the EO), outlined standards for safer and more responsible AI.[307] The EO was rescinded by President Donald Trump in January 2025, but it is nonetheless insightful to our discussion.[308]

Among other fields, the EO addressed human rights concerns, including bias and discrimination. In this context, the EO aimed, for instance, to promote fairness in the criminal justice system, and to guide different stakeholders, including landlords, federal contractors, and federal benefits programs, to prevent AI technologies from exacerbating discrimination.[309] However, notwithstanding their importance, the EO's provisions did not explicitly address linguistic concerns. Such concerns were also absent from the rest of the document, including the parts relating to the use of AI in healthcare, education, and work, even though the technologies' discrepancies across languages may have a dramatic influence in these fields.[310]

The EO's lack of attention to the linguistic aspects of AI might have stemmed, among other reasons, from its adoption procedure. The EO was a governmental regulation mechanism issued unilaterally by the US

---

[305] *See* WHITE HOUSE, *FACT SHEET: Biden-Harris Administration Executive Order Directs DHS to Lead the Responsible Development of Artificial Intelligence* (Oct. 30, 2023), https://www.dhs.gov/archive/news/2023/10/30/fact-sheet-biden-harris-administration-executive-order-directs-dhs-lead-responsible [https://perma.cc/998R-52KD] (concerning AI in general).

[306] Rebecca Fannin, *In Tech Rivalry with the US, China Is behind on a Key Asset: Its Own OpenAI,* CNBC (Mar. 31, 2024), https://www.cnbc.com/2024/03/31/in-ai-race-with-us-china-is-behind-on-a-key-weapon-its-own-openai.html [https://perma.cc/7TRM-MUEE].

[307] Executive Order 14110 of Oct. 30, 2023 (Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence).

[308] *Initial Rescissions of Harmful Executive Orders and Actions*, THE WHITE HOUSE (Jan. 20, 2025), https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/ [https://perma.cc/E85F-GM9E].

[309] Manuel Wörsdörfer, *Biden's Executive Order on AI: Strengths, Weaknesses, and Possible Reform Steps, AI and Ethics* (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4874596 [https://perma.cc/9N6G-ZBYS]. Moreover, the EO states that the administration "cannot—and will not—tolerate the use of AI to disadvantage those who are already too often denied equal opportunity and justice." *See* WHITE HOUSE, *supra* note 305, at sec. 2.

[310] *See* WHITE HOUSE, *supra* note 305.

https://doi.org/10.59015/wilj.SCTW4297

president, rather than a parliament-enacted legislation.[311] As such, the EO adoption process did not secure robust channels for DMLs (and other vulnerable sectors) to actively participate in drafting, negotiating, or adjusting this binding document. Moreover, even if these channels for influencing the document were provided, they may not have facilitated the representation of "all affected" people, per Fraser's theoretical frame, as the affected reference group is much wider than just American DML speakers. Given the key role American corporations play in shaping the global AI landscape, the EO lacked participative channels through which the linguistic needs of DMLs beyond the country could be debated and considered. Such channels are particularly important since the US and the American people are the primary reference groups for most of the EO. Even the EO's section dedicated to global cooperation did not mention issues of linguistic diversity and linguistic discrimination.[312]

AI is also regulated through regional and international legal tools. However, like the American EO, these tools, including the recently adopted Council of Europe Framework Convention on AI ("the AI Treaty") fail to directly address digital linguistic gaps.[313]

The AI treaty opened for the European member and non-member states' signatures in September 2024 and awaits ratification.[314] It was drafted by the Committee on Artificial Intelligence (CAI), an intergovernmental body encompassing the forty-six Council of Europe member states, eleven non-member states, and the European Union.[315]

---

[311] KENNETH R. MAYER, *Why Are Executive Orders Important?, in* WITH THE STROKE OF A PEN: EXECUTIVE ORDERS AND PRESIDENTIAL POWER 3, 4 (2001) ("Executive orders are, loosely speaking, presidential directives that require or authorize some action within the executive branch though they often extend far beyond the government. They are presidential edicts, legal instruments that create or modify laws, procedures, and policy by fiat"); *see also* BUREAU JUST. ASSISTANCE, *Executive Orders on Privacy and Civil Liberties and the Information Sharing Environment*, https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/executive-orders [https://perma.cc/NT9Y-2Z79] (last visited Dec 4, 2024).

[312] *See* WHITE HOUSE, *supra* note 305.

[313] *See* Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, May 9, 2024, C.E.T.S. No. 225.

[314] *Council of Europe Opens First Ever Global Treaty on AI for Signature*, COUNCIL EUR. NEWSROOM (September 5, 2024), https://www.coe.int/en/web/portal/-/council-of-europe-opens-first-ever-global-treaty-on-ai-for-signature [https://perma.cc/WJ2T-MDGA]; *see also*, *Committee on Artificial Intelligence (CAI)*, COUNCIL EUR., https://www.coe.int/en/web/artificial-intelligence/cai [https://perma.cc/MKX5-KVUR] (last accessed Jun. 10, 2024).

[315] Committee of Ministers, *Council of Europe Adopts First International Treaty on Artificial Intelligence*, COUNCIL EUR. NEWSROOM, https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence [https://perma.cc/S2J8-2X5H] (last accessed Jun 10, 2024). The non-member states include Argentina, Australia, Canada, Costa Rica, the Holy See, Israel, Japan, Mexico, Peru, the US, and Uruguay. *Id.*

https://doi.org/10.59015/wilj.SCTW4297

Additional stakeholders participated in the CAI discussions, including human rights experts and activists from governmental, international, and regional bodies, civil society, and the private sector.[316] This variety is certainly encouraging and might support the inclusion of vulnerable groups' voices in the drafting process of the document. Nonetheless, language-centered organizations or other designated representatives did not seem to partake in the CAI meetings, potentially explaining the absence of this pressing issue from the treaty.[317]

With no state or international governing framework to manage and guide digital linguistic opportunities and risks, digital platforms are left to make the call themselves. This is often a space where DML speakers cannot meaningfully engage in the decision-making processes either. One reason is the lack of built-in, accessible grievance mechanisms through which DML speakers (and other vulnerable groups) can address AI companies with concerns and requests regarding linguistic performance and bias.[318] Variations of such grievance mechanisms are currently provided to users of digital platforms such as Facebook, Instagram, and X.[319]

IV. PARTICIPATION AND EQUALITY IN INTERNATIONAL LAW

Applying Fraser's framework of justice to the digital linguistic landscape proved valuable in unveiling and identifying how the participation deprivation of DML speakers manifests across domains and contexts. However, the need to address these asymmetries not only rests on fairness grounds but also on a legal foundation.

This legal basis can be found, first and foremost, in the right to equality, a key value underpinning Fraser's "Parity of Participation" framework.[320] Indeed, a large body of legal literature anchors participation

---

[316] *See* Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, May 9, 2024, C.E.T.S. No. 225; *see* COUNCIL EUR. COMMM. ON A.I., LIST OF PARTICIPANTS (2022) https://rm.coe.int/cai-2022-lp1-fin/1680a6d913 [https://perma.cc/LR9P-TWXK].

[317] *Id.* Advocates of inclusion and equality in general did participate.

[318] *See, e.g.*, OpenAI, *ChatGPT*, https://openai.com/chatgpt/ [https://perma.cc/4L25-SWBX] (last accessed Feb. 6, 2025) (does not include grievance mechanism).

[319] *See, e.g.*, Facebook, *Request Review of Removed Content*, META, https://www.facebook.com/help/contact/741298560151661?_rdr [https://perma.cc/C6YY-54FB] (last visited Aug. 7, 2024); Oversight Board, *How we do Our Work*, META, https://www.oversightboard.com/our-work/ [https://perma.cc/FPB2-592S] (last accessed Aug 9, 2024).

[320] *See supra* Part III.A.

https://doi.org/10.59015/wilj.SCTW4297

within the wider discourse of equality.[321] Additional rights could be harnessed to protect DML speakers' participation, including linguistic rights, which are highly relevant to our discussion.[322] Nonetheless, the right to equality can encompass a wider range of the contemporary and future manifestations of DML speakers' participation challenges, including those identified through an application of Fraser's framework.[323] Equality is also a fundamental human right recognized by many local legal regimes and constitutions.[324] It is enshrined in various international law documents, among which are the Universal Declaration on Human Rights,[325] the International Covenant on Civil and Political Rights,[326] and the International Covenant on Economic, Social, and Cultural Rights.[327] The latter two are legally binding for states that ratified them.[328] Together, these three resources comprise the International Bill of Rights, which includes some of the most endorsed human rights requirements.[329]

How is the right to equality described and protected in these international law resources? Article 2(1) of the International Covenant on Civil and Political Rights stipulates, for instance, that states will respect and ensure to all individuals within their territory and subject to their jurisdiction "the rights recognized in the present Covenant, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."[330] Article 25 states that "every citizen shall have the right and the opportunity," without distinctions based, inter alia, on language

---

[321] JONATHAN RIX ET AL., EQUALITY PARTICIPATION AND INCLUSION 1: DIVERSE PERSPECTIVES (Jonathan Rix et al. eds., 2nd ed. 2010); Steven Wheatley, *Non-Discrimination and Equality in the Right of Political Participation for Minorities*, 3 J. ETHNOPOLITICS & MINORITY ISSUES EUR., 1 (2002).

[322] LIONEL WEE, *Introduction*, *in* LANGUAGE WITHOUT RIGHTS 3, 3–4 (Lionel Wee ed., 2010); *see* Robert Dunbar, *Linguistic Human Rights in International Law*, *in* THE HANDBOOK OF LINGUISTIC HUMAN RIGHTS 25 (Tove Skutnabb-Kangas & Robert Phillipson eds., 2022); *see also* discussion *infra* note 331.

[323] *See supra* Part III.A.

[324] *See, e.g.*, Pricilla A. Lambert & Druscilla L. Scribner, *Why Constitutions Matter for Gender Equality*, *in* GENDER, CONSTITUTIONS, AND EQUALITY 15, 24 (2023).

[325] G.A. Res. 217 (III) A, Universal Declaration of Human Rights, art. 1–2, 7 (Dec. 10, 1948).

[326] G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights (Dec. 16, 1966).

[327] G.A. Res. 2200A (XXI), International Covenant on Economic, Social and Cultural Rights (Dec. 16, 1966).

[328] *International Bill of Human Rights: A Brief History, and the Two International Covenants*, UNITED NATIONS, https://www.ohchr.org/en/what-are-human-rights/international-bill-human-rights [https://perma.cc/55PG-D6BR] (last accessed July 10, 2024).

[329] *Id.*

[330] International Covenant on Civil and Political Rights, *supra* note 326, art. 2(1).

https://doi.org/10.59015/wilj.SCTW4297

or origin, "[t]o take part in the conduct of public affairs, directly or through freely chosen representatives," and "[t]o have access, on general terms of equality, to public service in his country."[331] Finally, Article 27 prescribes that: "In those States in which ethnic, religious or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practice their own religion, or to use their own language."[332]

Among other provisions, Article 3 of the International Covenant on Economic, Social, and Cultural Rights requires states to "ensure the equal right of men and women to the enjoyment of all economic, social and cultural rights set forth in the present Covenant," and Article 15 enshrines the right "[t]o take part in cultural life;" and "[t]o enjoy the benefits of scientific progress and its applications."[333]

As states are the immediate and first subject of international law, the International Covenants' provisions directly apply to and bind states that ratified them.[334] The right to equality, as enshrined in the International Bill of Rights, could thus serve as an apt legal foundation to oblige states to promote DML speakers' participation by investing the resources, time, and labor, as well as the regulatory and policy-making efforts needed for more equal distribution, recognition, and representation of and for DMLs.

While states play a pivotal role in shaping the digital linguistic landscape, private actors, particularly AI companies, are also central players in this context. What is, if any, the legal foundation for requiring AI companies to take action to facilitate DML speakers' participation?

The legal requirements of private actors concerning equality and human rights are more delicate and, often, more charged than those that apply to states. In the digital context though, the last decade has seen a

---

[331] *Id.*, art. 25.

[332] *Id.*, art. 27.

[333] International Covenant on Economic, Social and Cultural Rights, *supra* note 326, art. 3, 15. Linguistic equality has a central place in many other international law documents, some are very relevant to our discussion, though not necessarily legally binding. *E.g.* G.A. Res. 47/135, Declaration on the Rights of Persons Belonging to National or Ethnic, Religious, and Linguistic Minorities (Dec. 18, 1992); UNIVERSAL DECLARATION OF LINGUISTIC RIGHTS FOLLOW-UP COMMITTEE, UNIVERSAL DECLARATION OF LINGUISTIC RIGHTS 21–22 (1998); UNESCO, *Universal Declaration on Cultural Diversity* (Nov. 2, 2001) https://www.ohchr.org/en/instruments-mechanisms/instruments/universal-declaration-cultural-diversity [https://perma.cc/S5YB-CM8M].

[334] *International Bill of Human Rights*, *supra* note 326.

https://doi.org/10.59015/wilj.SCTW4297

proliferation of literature discussing private actors' far-reaching influence and the legal justifications for applying fairness, accountability, and human rights-related requirements to them.[335] In addition, self-imposed obligations,[336] industry codes of conduct,[337] and, increasingly, binding regulations, have emerged and influenced the lines of private actors' legal roles.[338]

A valuable source that could assist in calibrating our expectations of digital platforms and technology companies regarding digital linguistic equality and DML speakers' participation is the United Nations' Guiding Principles on Business and Human Rights (UNGPs). The UNGPs were developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises as a means of implementing the United Nations' "Protect, Respect and Remedy" Framework. They rely on the notion that private actors are significant players in shaping the human rights landscape.[339] The Human Rights Council endorsed the Guiding Principles in 2011.[340]

The "Protect, Respect and Remedy" Framework offers a balanced and delicate mechanism to address private actors' human rights role. It encompasses "the State duty to protect against human rights abuses," by such private actors, "through appropriate policies, regulation, and

---

[335] *See, e.g.*, Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018); NICHOLAS P. SUZOR, *The Hidden Rules of the Internet*, *in* LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 6, 6–9 (2019); TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (Yale Univ. Press, 2018); K. Sabeel Rahman, *The New Utilities: Private Power, Social Infrastructure, and the Revival of the Public Utility Concept*, 39 CARDOZO L. REV. 1621 (2018); Ira Steven Nathenson, *Super-Intermediaries, Code, Human Rights*, 8 INTERCULTURAL HUM. RTS. L. REV. 19, 158 (2013); Noa Mor, *No Longer Private: On Human Rights and the Public Social Network Sites*, 47 HOFSTRA L. REV. 651 (2018).

[336] See, e.g., discussion on the voluntary adoption of the UNGPs.

[337] *See, e.g.*, OECD.AI, *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*, (Nov. 14, 2023) https://oecd.ai/en/catalogue/tools/voluntary-code-of-conduct-on-the-responsible-development-and-management-of-advanced-generative-ai-systems [https://perma.cc/F2YH-MLZP].

[338] *See, e.g.*, *id.*

[339] UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, GUIDING PRINCIPLES ON BUSINESS AND HUMAN RIGHTS: IMPLEMENTING THE UNITED NATIONS "PROTECT, RESPECT AND REMEDY" FRAMEWORK 3 (2011). The guidelines were described in the Special Representative's final report to the Human Rights Council. *See* John Ruggie (Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises), *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, U.N. Doc. A/HRC/17/31 (March 21, 2011).

[340] UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, *supra* note 339, at iv.

https://doi.org/10.59015/wilj.SCTW4297

adjudication."[341] The United Nations framework also points out corporations' "responsibility to respect human rights," according to which, "business enterprises should act with due diligence to avoid infringing on the rights of others and to address adverse impacts with which they are involved."[342] The UNGPs' commentary clarifies:

> The responsibility to respect human rights is a global standard of expected conduct for all business enterprises wherever they operate. . .it exists over and above compliance with national laws and regulations protecting human rights. Addressing adverse human rights impacts requires taking adequate measures for their prevention, mitigation and, where appropriate, remediation.[343]

The guidelines emphasize that private actors' responsibility to respect human rights relates to internationally recognized human rights, that are "understood, at a minimum, as those expressed in the International Bill of Human Rights," and thus cover the right to equality as discussed above.[344]

The UNGPs enshrine private companies' responsibility to "[a]void causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur." The guidelines also cover private companies' responsibility to work towards preventing or mitigating hindrances to human rights caused by their operations, products, or services, regardless of the companies' contribution to those impacts.[345]

The UNGPs also refer to the responsibilities of business enterprises to have operational-level grievance mechanisms for individuals and groups adversely impacted by them.[346] The UNGPs' commentary further explains that these grievance systems

> support the identification of adverse human rights impacts. . .by providing a channel for those directly impacted by the enterprise's operations to raise concerns when they believe they are being or will be adversely impacted. By analyzing trends and patterns in complaints,

---

[341] Ruggie, *supra* note 337, at 4.

[342] *Id.* The third component of the Framework concerns "the need for greater access by victims to effective remedy, both judicial and non-judicial."

[343] UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, *supra* note 339, at 13.

[344] *Id.* at 12.

[345] *Id.* at 14.

[346] *Id.* at 31.

https://doi.org/10.59015/wilj.SCTW4297

> business enterprises can also identify systemic problems and adapt their practices accordingly.[347]

While not legally binding,[348] the UNGPs have been adopted by various digital corporations, including Meta,[349] Microsoft,[350] and Apple.[351] They are widely endorsed as the threshold that private actors should meet.[352] Even as a "soft law" resource, they contribute to the substantive message regarding AI companies' responsibilities in facilitating DML speakers' participation. Furthermore, their in-depth content offers a practical starting point for articulating these companies' responsibilities, as will be later demonstrated.

Finally, the UNGPs distinguish between different private actors, thereby offering a flexible and balanced approach. The UNGPs clarify that:

> [t]he responsibility of business enterprises to respect human rights applies to all enterprises regardless of their size, sector, operational context, ownership and structure.[353] Nevertheless, the scale and complexity of the means through which enterprises meet that responsibility may vary according to these factors and with the severity of the enterprise's adverse human rights impacts.[354]

The UNGPs are also useful in the digital context. The B-Tech Project, launched by the United Nations in 2019, focuses on the UNGPs'

---

[347] *Id.* at 32.

[348] *See generally* Ionel Zamfir, *Towards a Binding Treat on Business and Human Rights: Despite Progress, Still no Final Outcome in View*, (May 2022) https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2022)729435 [https://perma.cc/F66E-XFR8] (For the efforts to create a binding treaty on human rights and business within the UN and European Union frameworks).

[349] *Corporate Human Rights Policy*, META, https://about.fb.com/wp-content/uploads/2021/04/Facebooks-Corporate-Human-Rights-Policy.pdf [https://perma.cc/T5TD-6R93] (last visited Aug 9, 2024).

[350] Steve Crown, *Taking on Human Rights Due Diligence*, MICROSOFT ON THE ISSUES (Oct. 20, 2021), https://blogs.microsoft.com/on-the-issues/2021/10/20/taking-on-human-rights-due-diligence/ [https://perma.cc/VU7G-UZAD].

[351] *Our Commitment to Human Rights*, APPLE (May 2024), https://s2.q4cdn.com/470004039/files/doc_downloads/gov_docs/Apple-Human-Rights-Policy.pdf [https://perma.cc/QDM2-E93H].

[352] *The UN Guiding Principles in the Age of Technology*, UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R (Sept. 2020), https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/introduction-ungp-age-technology.pdf [https://perma.cc/Z3G6-5W4G].

[353] UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, *supra* note 339, at 15.

[354] *Id.*

https://doi.org/10.59015/wilj.SCTW4297

implementation in digital settings,[355] including in relation to Generative AI companies.[356] Among other impacted human rights, the project identifies Generative AI's potential risks regarding the right to equality, specifically warning that:

> "Low resource languages" are often underrepresented in generative AI training datasets, leading to underperformance of generative AI systems for speakers of these languages. Underperformance of generative AI for users from certain linguistic, geographic, and cultural backgrounds may in itself constitute a form of discrimination, and threatens to widen the growing digital divide between high-resource and low-resource countries.[357]

To conclude, international law offers the legal foundation for requiring states to take actions to reduce linguistic hierarchies and facilitate participation for DML speakers. It can also assist us in setting our expectations of private actors in this regard. In the next and final Part, I detail how this international law basis could be harnessed to further DML speakers' participation, as understood in Fraser's "Parity of Participation" framework.

V. TYING TOGETHER TECHNOLOGY, JUSTICE, AND LAW: TOWARDS PARTICIPATION OF DML SPEAKERS IN DIGITAL DOMAINS

In the previous Parts, I have discussed the digital linguistic gaps and explored the techno-social causes driving them in the NLP and LLMs context. Drawing on Nancy Fraser's work, I examined how these disparities systematically limit DML speakers' ability to participate in digital domains. I then discussed the international law foundation for requiring states and private actors to facilitate such participation.

---

[355] *B-Tech Project*, UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, https://www.ohchr.org/en/business-and-human-rights/b-tech-project [https://perma.cc/3BX7-HEG9] (last visited July 13, 2024).

[356] *See also Generative AI Human Rights Due Diligence Project UN Human Rights,* UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R (May 2023), https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/B-Tech-Generative-AI-concept-note.pdf [https://perma.cc/U4VP-PQFK].

[357] *Taxonomy of Human Rights Risks Connected to Generative AI,* UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf [https://perma.cc/VE4T-3LZY] (last visited Aug. 9, 2024).

https://doi.org/10.59015/wilj.SCTW4297

I will now tie together the technological insights, legal discussion, and justice-based participation paradigm, offering steps to mitigate DML speakers' current distribution, recognition, and representation barriers. These suggestions will address the roles of both states and private companies. They may, of course, be adjusted according to the circumstances and contexts at hand, including the financial resources of governments and private bodies and their unique influence on DML speakers' participation and linguistic diversity.[358]

A. TOWARDS EQUAL DISTRIBUTION

As discussed above, maldistribution concerns the asymmetric allocation of resources, opportunities, and freedoms. It is the result of economic structures that support and secure these gaps. Moving towards equal distribution necessitates, thus, an organized and systematic intervention that targets these adverse economic structures.[359]

In the context of LLMs, the root causes of the linguistic disparities include the training data and training processes on the one hand and design and evaluation choices and constraints, on the other hand.[360] How should states and private companies tackle these challenges?

From the states' end, efforts should be directed to meaningfully diversify both labeled and unlabeled datasets and represent DML content in LLM training. Creating unlabeled datasets might require seeking new sources and channels for such corpora and creating designated business models to support them. This may involve, inter alia, responsibly leveraging the data created and maintained by DML communities as well as their public bodies and civil organizations and, if appropriate, compensating for it.[361] Diversifying unlabeled datasets may include efforts to convert data from one form to another (for instance, by using speech-to-text technologies)[362] and, when feasible, to encourage creating content from scratch. To meaningfully diversify labeled datasets, it may be required to rely on and encourage linguistic research and invest in human

---

[358] See *supra* Part IV explaining the UNGPs' nuanced approach in this regard.

[359] *See supra* Part III.

[360] *See supra* Part II.B.2.

[361] This will require caution concerning private data usage and verifying that DML speakers' vulnerability is not exploited.

[362] *See* Wushour Slam et al., *Frontier Research on Low-Resource Speech Recognition Technology,* SENSORS, Nov., 2023 at 1, 33.

https://doi.org/10.59015/wilj.SCTW4297

and (responsible) automatic annotation.[363] Unlabeled and labeled datasets should be made as interoperable and open as possible to reduce access barriers.[364] To enhance LLM performance in DMLs, investments should be directed at finding innovative methods to improve zero-, one-, and few-shot learning approaches—which benefit DMLs through the reliance on the abundant DDL data—as these learning approaches' capabilities are currently lacking in DMLs.[365] However, noting the fundamental cultural importance of original DML training data, such approaches should not negate parallel efforts to diversify training datasets and processes.[366] Creating and improving DML high-quality datasets should, in turn, incentivize LLM developers to include DML data in their language-specific or multilingual models, enlarge the amount of DML content within the overall pretraining sets, and integrate DML examples and instructions in the fine-tuning stages. Nonetheless, direct interventions should be independently directed towards creating such incentives as well.

States should also invest resources into adjusting LLMs' design and evaluation processes. They should critically assess and rethink DML tokenization processes—some of which are currently skewed toward DDLs—to enhance linguistic fairness and equality.[367] This may involve, as discussed above, leveraging designated tokenizers for DMLs.[368] The processes used to filter datasets used for training should be carefully calibrated to not disproportionately censor vulnerable linguistic groups, including minorities and speakers of nonstandard English.[369] Such diligence might be achieved by combining different filtering approaches, for instance.[370] In addition, to provide a more reliable picture of LLM performance across languages, states should work towards facilitating evaluation processes in DMLs and not only in English and other DDLs. Finally, states should encourage and support the creation of benchmarks in a wide range of DMLs to better assess multilingual abilities and to reduce bias and mistakes inserted into these thresholds through translation.

Furthermore, states' obligations may include setting standards, policies, and regulations to support linguistic diversity and establish

---

[363] *See supra* Part II.A, II.B.2.
[364] *Id.*
[365] *See supra* Part II.B.2.
[366] *See supra* Parts I.A, II.B.2.
[367] *See supra* Part II.B.2.
[368] *Id.*
[369] *Id.*
[370] *Id.*

https://doi.org/10.59015/wilj.SCTW4297

channels for collaboration among industry, research, and state bodies. It should be noted that states' obligations to support DML speakers' participation are further underscored by governments' increasing reliance on AI tools and the potential harms they bring. Examples of such harms are erroneous sanctions that DML speakers may face due to disparities in AI tools' performance.[371]

For private bodies, the requirement to facilitate participation should be more moderate in comparison to states but nonetheless valid.[372] Applying the UNGPs may indicate AI companies' responsibility to create and set relevant policies, including those that reflect a corporate commitment to linguistic diversity and to identifying and mitigating linguistic disparities.[373] The policies should be "informed by relevant internal and/or external expertise," stipulate "the enterprise's human rights expectations of personnel, business partners and other parties directly linked to its operations, products or services," and be reflected in operational protocols and procedures "necessary to embed it throughout the business enterprise."[374] The UNGPs also anchor private actors' responsibility to perform relevant human rights due diligence and impact assessments and to tackle the linguistic gaps that they find.[375] Private companies should also, according to the UNGPs, track whether they are adequately addressing adverse human rights impacts. Such tracking should be based "on appropriate qualitative and quantitative indicators," and "[d]raw on feedback from both internal and external sources, including affected stakeholders."[376]

Of course, as discussed above, the UNGPs acknowledge that the nature of their application by private bodies should vary according to factors such as these bodies' size and ownership.[377] Dominant AI companies, or those acquired by large technology corporations like Google (which has developed and owns Gemini, along with additional LLMs) should, therefore, meet a higher bar of responsibility than less dominant AI private actors whose market share and financial capabilities are smaller.

---

[371] *See supra* Part III.B.1.

[372] *See supra* Part IV.

[373] UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, *supra* note 339, at 15.

[374] *Id.* at 16.

[375] *Id.* at 17–20.

[376] *Id.* at 22. Providing users with grievance mechanisms could be one such resource. See discussion *supra* Part III.B.3 and *see* discussion *infra* Part V.B.

[377] UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, *supra* note 339, at 15.

https://doi.org/10.59015/wilj.SCTW4297

B. TOWARDS RECOGNITION

Fostering recognition of DML speakers requires creating channels to evaluate and address the institutionalized, disfavored cultural value ascribed to them in digital domains. Recognizing DML individuals and communities in the LLM landscape involves creating sustainable channels through which their voices and needs become visible, present, and acknowledged. To be sure, extensively covering the intricate set of considerations relevant to this issue—particularly against the backdrop of the long-lasting linguistic oppression processes discussed above[378]—exceeds the scope of this Article. I will, therefore, only briefly offer some central directions that should be pursued in the LLM context.

As arises from the techno-social discussion concerning LLMs' training and design drawbacks, promoting recognition should be based on the understanding that excluding DML data hampers and cements the disfavored cultural value ascribed to it.[379] This also applies to DDL content translated to DMLs, since it may fail to carry the latter's unique narratives, values, and needs.[380] Such translated content also involves "assimilation to majority or dominant cultural norms," a state that misaligns with Nancy Fraser's premise.[381]

For states, fostering DML speakers' recognition may involve dedicating resources to raising awareness among industry, research, and government sectors concerning digital linguistic gaps and their implications. This may also require collaboration with advocacy organizations dedicated to furthering vulnerable communities' rights or linguistic and cultural diversity. Supporting DML speakers' recognition may also entail other measures, such as creating opportunities for computer science students and professionals to learn about these disparities and how to address them. It may also necessitate establishing communal, governmental, and international bodies or task forces, and regulatory measures that apply to public and private bodies.

Private bodies' role in facilitating DML speakers' recognition should include the creation of transparency and accountability channels through which the causes and manifestations of linguistic biases in NLP technologies will be discussed and addressed. Among other needed

---

[378] *See supra* Part I.B.

[379] *See supra* Parts II.B.2, III.B.2.

[380] *See supra* Part III.A.1.

[381] Fraser, *supra* note 267, at 21.

https://doi.org/10.59015/wilj.SCTW4297

mechanisms, this can be achieved by submitting periodic transparency reports which cover various aspects of these private bodies' performance, as discussed earlier.[382] Private companies should also work towards more recognition of DML speakers by creating engagement mechanisms with various stakeholders, including DML communities, and by emphasizing and demonstrating dedication to linguistic diversity, both internally and externally. Private companies' responsibilities may also encompass impact assessments of their existing and in-development products, to detect cultural biases against DML speakers.

C. TOWARDS REPRESENTATION

Following distribution and recognition, the final domain for facilitating DML speakers' participation concerns their ability to influence decision-making processes that impact them. Partnership in the decision-making process provides avenues through which DML speakers can draw attention to linguistic disparities in digital domains and take action to address them.[383] This is a political measure necessary to tackle the systemic sidelining of DMLs.[384]

For states, tackling misrepresentation may require rethinking regulatory and policymaking mechanisms, as well as the circle of stakeholders that can influence such decisions. This may entail preferring legislative mechanisms that rely on broad and open deliberation processes, and secure spaces for DML speakers to influence and be heard. The adoption process of the Council of Europe treaty on AI, mentioned above, introduces a positive direction that could have been further improved had designated DML representatives participated. On the other hand, governmental regulation, such as the US EO, lacks spaces for DML speakers (and other vulnerable groups) to be adequately regarded.[385]

Moreover, given that the influence of American AI companies on DML speakers extends far beyond their territory, DML speakers should be heard when formulating regulations that apply to these companies. This responsible approach aligns with Fraser's notion that "all those affected

---

[382] *See supra* Part III.B.2.
[383] *See supra* Parts III.A, III.B.3.
[384] *See supra* Parts III.A, III.B.3.
[385] *See supra* Part III.B.3.

https://doi.org/10.59015/wilj.SCTW4297

by a given social structure of institution" should be included in the decision-making process shaping it.[386]

Of course, representation is not limited to the drafting of regulations and may encompass a wide tapestry of channels through which DML speakers can be heard and assume an active role in crafting their future digital lives. Examples may include inviting DML speakers and organizations to relevant discussions within governmental departments, soliciting their feedback on pressing issues, and appointing them to designated task forces.[387]

As for private AI companies, they may contribute to DML speakers' representation by seeking their DML users' feedback on existing and future products and diversifying their own teams. Private bodies can also establish grievance mechanisms that allow DML users to point out linguistic gaps and concerns, as highlighted in the UNGPs.[388] Beyond empowering DML speakers, analyzing such feedback can enable AI companies to more effectively address linguistic disparities.[389]

VI. CONCLUSION

Digital participation dramatically varies among speakers of different languages. Of the world's seven thousand languages, only speakers of select dominant languages can fully enjoy the far-reaching advantages digital avenues afford. Speakers of the remaining languages have limited access to these avenues or are excluded altogether. These linguistic asymmetries, rooted in long-standing processes of dominance and oppression, manifest in poor connectivity, inadequate equipment, and unavailable applications and services. The emergence of NLP and LLMs further exacerbates and solidifies these inequalities due to biases in training data and training processes, design choices, evaluation approaches, and benchmarks. Drawing on Nancy Fraser's "Parity of Participation" trifold framework and on the international law human rights

---

386 FRASER, *supra* note 37, at 24.

387 Sherry R. Arnstein, *A Ladder of Citizen Participation*, 35 J. AM. INST. PLANNERS 216, 219 (1969) (discussing a wide range of participation possibilities); Archon Fung, *Varieties of Participation in Complex Governance,* 66 PUBLIC ADMIN. REV. 66, 68 (2006).

388 *See supra* Part IV.

389 Nick Devas & Ursula Grant, *Local Government Decision-Making – Citizen Participation and Local Accountability: Some Evidence From Kenya and Uganda*, 23 PUBLIC ADMIN. AND DEV. 307, 308 (2003); Michel Pimbert & Tom Wakeford, *Overview – Deliberative Democracy and Citizen Empowerment,* 40 PLA NOTES 23, 23 (2001); CAROLE PATEMAN, PARTICIPATION AND DEMOCRATIC THEORY 42 (Cambridge Univ. Press, 1970).

https://doi.org/10.59015/wilj.SCTW4297

settings, this Article analyzes the nature and implications of this pressing linguistic sidelining and outlines a way forward, spanning technological, economic, cultural, and regulatory considerations.

https://doi.org/10.59015/wilj.SCTW4297