Decolonizing Linguistic Policies in Speech AI: A Framework for Cross-Culturally Competent Systems

Anonymous submission to Interspeech 2026

Abstract

Speech AI increasingly mediates access to public services, healthcare, and education, yet it routinely fails speakers of low-resource, Indigenous, and non-standard language varieties. We argue these failures are not merely technical but reflect implicit linguistic policies that reproduce colonial language hierarchies. Drawing on Bourdieu's linguistic capital, racio-linguistic ideology, and decolonial computing, we show how choices in training data, evaluation metrics, and language model priors operate as political acts that determine whose voices become machine-legible. We introduce the Three Harms (3M) taxonomy—Misrecognition, Misalignment, and Mistrust—to capture failure modes beyond word error rate. We then propose a Participatory Framework for Culturally Competent Speech AI: participatory auditing, community co-design, equitable deployment, and feedback integration, positioning Global South communities as co-designers rather than passive targets of inclusion.

Index Terms: speech recognition, linguistic bias, de-colonial AI, low-resource languages, participatory design, cultural competence

1. Introduction

Speech AI systems now mediate access to public services, healthcare, education, and legal processes, yet they routinely fail speakers of low-resource, indigenous, and non-standard varieties [1, 2]. These failures are not isolated bugs: they are the predictable outcomes of how systems define which voices are legible to machines, with material consequences ranging from denied access to the indignity of being rendered unintelligible [3, 4]. We argue that the core issue is structural, not incidental.

We term these structural decisions **linguistic policies**: the explicit and implicit design rules that govern (i) *data* (which languages, dialects, and speaking contexts are collected and labeled), (ii) *metrics* (what counts as "correct" and how error is weighted across speakers), and (iii) *priors* (the language-model assumptions that privilege certain speech patterns). These policies reproduce colonial hierarchies of linguistic value by embedding prestige norms into the pipeline [5, 6].

This paper makes four contributions:

- We provide a theoretical account of linguistic policies in speech AI, synthesizing linguistic capital theory, racio-linguistic ideology, and de-colonial computing to explain why exclusion is produced by design choices rather than mere data scarcity [7, 8, 9, 10, 11, 12].
- We introduce a Seven-layer Situatedness Model for Linguistic Diversity in Speech AI. Moving upward adds socio-technical context; colonial linguistic hierarchies shape what is treated as "standard," "supported," and "valuable.
- We introduce the Three Harms (3M) taxonomy: Misrecognition, Misalignment, and Mistrust, to capture failure modes that are invisible to WER-focused evaluation.
- We formalize a Participatory Framework for Culturally Competent Speech AI that reassigns design authority to affected language communities.
- We connect community-led initiatives across the Global South to a replicable design methodology that operationalizes cultural competence in speech AI.

2. Theoretical Foundations

Our analysis integrates three intellectual traditions that, together, illuminate the political architecture of speech AI: linguistic capital theory, racio-linguistic ideology, and de-colonial computing.

2.1. Linguistic Capital and Market Value

Pierre Bourdieu's concept of *linguistic capital* provides a structural account of why certain language varieties are treated as more valuable than others [7]. For Bourdieu, language operates within a "linguistic market" where utterances are evaluated not solely by their communicative content but by their symbolic alignment with dominant norms. Speakers of high-resources varieties such as Standard American English, Metropolitan French, and European Portuguese or Peninsular Spanish as colonial reference standards, accumulate linguistic capital that converts into social, economic, and institutional power. Speakers of marginalized varieties (e.g., Brazilian Portuguese in its regional and Afro-Brazilian forms, or Latin American Spanish in its diverse national (Mexican, Dominican, Colombian) and Indigenous-contact varieties) face a "symbolic tax": their speech is devalued, misheard, or ignored [13, 14].

Speech AI systems operationalize Bourdieu's linguistic market in algorithmic form. When an ASR system achieves a 5% WER for Standard American English but a 35% WER for African American Language [3], or when a voice assistant supports Parisian French but not Senegalese French, it is not merely reflecting data availability, it is *reproducing* the market valuation of these language varieties. The training data itself encodes market position: languages with high institutional power (English, Mandarin, Standard Spanish) are overrepresented in digital corpora, while languages with lower institutional power (Wolof, Igbo, Akan) remain "low-resource"—a term that obscures the political economies of neglect that produced the resource gap [15].

How does a language become "low-resource"? To avoid treating "low-resource" as a monolithic label, we draw on Joshi et al.'s six-class taxonomy of linguistic inclusion, which partitions the world's languages by their digital and NLP resource profiles [16]. The taxonomy ranges from **(0) Left-Behinds** (e.g., Dahalo, Warlpiri, Popoloca, Wallisian, Bora) through **(1) Scraping-Bys** (e.g., Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo) and **(2) Hopefuls** (e.g., Zulu, Konkani, Lao, Maltese, Irish), to **(3) Rising Stars** (e.g., Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew), **(4) Underdogs** (e.g., Russian, Hungarian, Vietnamese, Dutch, Korean), and **(5) Winners** (e.g., English, Spanish, German, Japanese, French) [16]. We use this breakdown as a descriptive measure of how resource allocation and research attention are distributed, and as motivation for why culturally competent speech AI must address not only data scarcity but also the political and institutional conditions that determine which language varieties become "supported" in the first place. Importantly, our claim is that speech AI systems often inherit the value judgments embedded in these resource tiers, treating "Winners" as normative and pushing culturally situated speech patterns in other classes into misrecognition, misalignment, and mistrust.

2.2. Racio-linguistic Ideologies in Technology

Rosa and Flores's *racio-linguistic* framework extends this analysis by demonstrating how perceptions of language are inextricable from perceptions of racialized bodies [8]. The "white listening subject"—their term for the dominant perceptual stance that positions white, middle-class language practices as neutral—does not merely hear language differently; it *construes* the speech of racialized speakers as deficient regardless of its linguistic content [9]. This is not a failure of listening accuracy but an ideological orientation that treats some speakers as inherently less intelligible.

We propose that speech AI systems function as *algorithmic listening subjects* that encode racio-linguistic ideologies at scale. When ASR systems are trained on datasets that disproportionately represent standard varieties, when pronunciation models are benchmarked against prestige norms, and when error tolerance thresholds are calibrated to dominant-group performance, the resulting system "listens" through the same ideological lens that Flores and Rosa describe [9]. The system does not hear what is spoken; it hears what it has been designed to recognize as speech. Lippi-Green's foundational work on linguistic discrimination demonstrates how such ideological orientations operate in institutions from courts to classrooms [17]; speech AI extends this discrimination into automated infrastructure.

2.3. De-colonial Computing

De-colonial computing provides the third pillar. Ali frames de-colonial computing as a critical orientation that interrogates how coloniality—the enduring structures of colonial power—persists within computational systems, epistemologies, and design practices [10]. Irani et al.'s postcolonial computing lens reveals how technology design naturalizes Western epistemologies while rendering other knowledge systems invisible or deficient [11]. Dourish and Mainwaring expose the "colonial impulse" in ubiquitous computing, where systems designed for Western contexts are deployed globally with minimal adaptation [18].

Mohamed et al. synthesize these insights into a framework for de-colonial AI, identifying algorithmic exploitation, algorithmic dispossession, and the imposition of dominant epistemologies as key mechanisms through which AI systems perpetuate colonial relations [12]. We build on this framework to argue that speech AI represents a particularly acute site of colonial continuity: language was a primary instrument of colonial governance [5, 19, 6], and speech AI systems that privilege colonial languages while marginalizing indigenous ones extend this governance into digital infrastructure.

2.4. The Digital Language Divide

These theoretical frameworks converge in what Joshi et al. document as the "digital language divide" [4]. Their taxonomy classifies the world's languages into categories from "The Winners" (a handful of languages with extensive NLP resources) to "The Left-Behinds" (thousands of languages with virtually no digital presence). Blasi et al. demonstrate that language technology performance correlates not with linguistic complexity but with socioeconomic power: the languages that perform worst in NLP systems are spoken by the world's most economically marginalized populations [20]. This is not a coincidence; it is, as Couldry and Mejias argue, a form of *data colonialism* in which the extraction of value from data reproduces colonial patterns of dispossession [21].

The convergence of linguistic capital theory, racio-linguistic ideology, and de-colonial computing reveals that speech AI's exclusions are not merely the product of "insufficient data" awaiting technical remediation. They are the algorithmic expression of centuries-old hierarchies of linguistic and racial value. Addressing them requires not only more data but fundamentally different relationships between speech systems and the communities they purport to serve.

To make this layered structure explicit, we introduce a seven-layer situatedness model (Figure 1) that links language, nation, region, ethno-linguistic identity, ideology, justice, and socio-technical implications. This model clarifies why "supporting a language" is never a single technical task: it requires attending to layered social and political conditions that shape intelligibility and harm.

3. Positionality Statement

In alignment with Haraway's account of *situated knowledges* and the privilege of partial perspective [22] and Suchman's formulation of *located accountabilities* in technology production [23], we view this paper as shaped by our standpoints as researchers whose identities, linguistic communities, and institutional locations condition what we notice as harm, what we treat as evidence, and what we imagine as responsible intervention. We do not claim neutrality: our analysis of "linguistic policies" in speech AI is informed by the ways language hierarchies are lived and enforced through schooling, institutions, markets, and everyday interactions, and by our recognition that speech and language technologies can reproduce these hierarchies when they privilege prestige norms.

Our author team brings lived and scholarly experience across multiple language communities that sit in different relations to colonial and postcolonial language power, including African American English (AAVE) and other Afro-diasporic English varieties of North America and the Carribean, Brazilian Portuguese and its social/ethnic variations plus indigenous language use, Nigerian-English plus Yorùbá and Ìgbò tribal languages, Mexican/Latin American Spanish varieties, and Ghanaian-English plus Twi and Akan tribal languages. This range of linguistic and cultural standpoints is central to the paper's argument that cultural competence cannot be reduced to language-level "coverage" or aggregate accuracy. It also in-

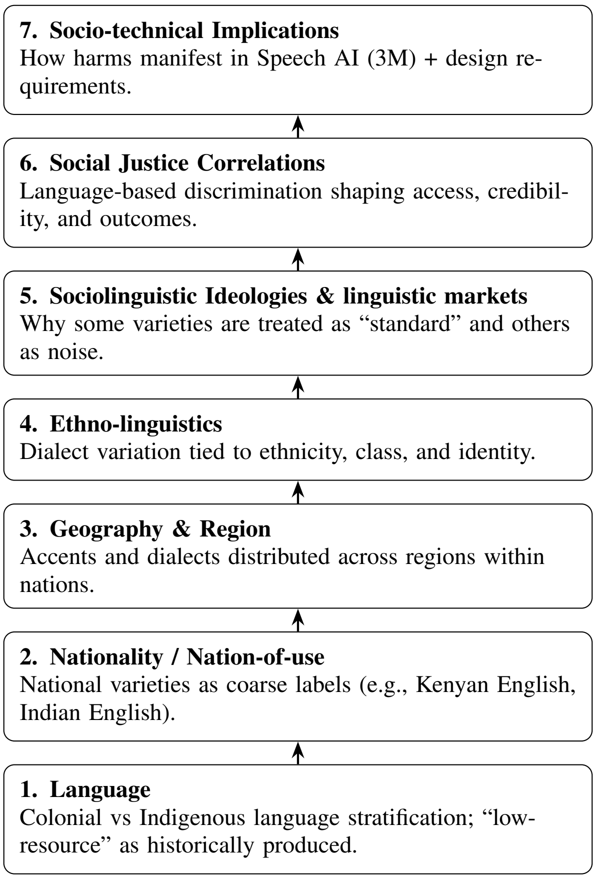

Figure 1: *A seven-layer situatedness model for linguistic diversity in speech AI. Moving upward adds socio-technical context; colonial linguistic hierarchies shape what is treated as "standard," "supported," and "valuable."*

forms our emphasis on harms that exceed transcription error, including pragmatic and cultural misalignment and the downstream mistrust that follows from repeated system failure.

Our standpoints also shape our methodological commitments. First, our approach prioritizes community-led and regionally grounded language technology efforts (e.g., pan-African and local NLP initiatives, and Global South research networks) as sites of expertise rather than peripheral "data sources." Second, our proposed participatory framework reflects a normative stance: affected language communities should be treated as co-designers and co-auditors of speech AI systems, with meaningful influence over evaluation criteria, deployment conditions, and pathways for repair and redress. Finally, we recognize that positionality is not resolved through disclosure alone. We therefore treat this statement as an accountability mechanism: a reminder that our claims are partial, that our recommendations must remain responsive to community critique, and that the work ahead requires sustained collaboration beyond what any single paper can represent.

We also acknowledge that we write from university institutions in North America, which shape access to publication venues, resources, and legitimacy; this paper therefore aims to amplify and remain accountable to Global South knowledge-making rather than treat it as supplementary.

Table 1: *How to spot linguistic policies in speech AI: sites of decision-making and observable signals.*

Policy site	**Concrete signals**
Training data	Benchmark composition skews toward prestige languages; crowdsourcing excludes rural or low-connectivity speakers.
Training data	Transcription guidelines enforce standardized orthography over local spellings or code-switching.
Metrics	Primary evaluation uses WER without domain- or community-specific weighting.
Metrics	"Ground truth" assumes a single correct transcript despite multiple legitimate variants.
Model priors	Language model favors monolingual sequences; code-switching is penalized.
Model priors	Normalization rules remove honorifics or politeness markers as "noise."
Deployment	Refusal or fallback behavior defaults to English-only responses for unsupported varieties.

4. Linguistic Policies as Design Choices

We define **linguistic policies in speech AI** as the explicit and implicit design decisions that govern which language varieties a system supports, how it evaluates correctness, and what constitutes intelligible speech. These policies operate at three interconnected levels: data, metrics, and modeling.

Table 1 provides a compact diagnostic for identifying where linguistic policies are enacted in practice and how they surface in system behavior. This diagnostic is not exhaustive, but it makes policy sites visible and actionable for audit and redesign.

4.1. Training Data Curation as Policy

The selection, sourcing, and annotation of training data constitute the most consequential linguistic policy decisions in speech AI. When developers prioritize publicly available speech corpora, which overwhelmingly represent English, Mandarin, and a small number of European languages [4], they enact a policy of linguistic triage that mirrors colonial hierarchies. The term "low-resource language" obscures the political economy of this resource distribution: these languages are not inherently resource-poor; they have been *made* resource-poor through centuries of institutional marginalization [15].

Consider the historical production of this "scarcity." English was imposed upon ethnic Africans in Ghana, Nigeria, Kenya, and South Africa—nations that each possess rich arrays of native languages such as Akan, Ga, Swahili, and Zulu. Spanish was forced upon Indigenous Americans and enslaved Africans in the Caribbean. Portuguese was imposed upon Indigenous and Afro-Brazilian populations. French was enforced across West African nations including Senegal and Ivory Coast, and in Haiti [5, 6]. These colonial languages subsequently became the languages of digital infrastructure, institutional documentation, and academic publishing, which are the very sources from which training corpora are derived. When a speech AI system is "trained on available data," it is trained on the documentary legacy of colonial language dominance.

4.2. Evaluation Metrics as Policy

Word Error Rate (WER) remains the dominant metric for ASR evaluation, yet its use as a primary measure of system quality enacts its own linguistic policy. WER treats all word-level errors as equivalent, irrespective of their communicative or social consequences. It cannot capture whether a transcription error changes the meaning of an utterance, whether it renders a speaker's intent unintelligible, or whether it has differential consequences in high-stakes contexts such as healthcare or legal proceedings [24].

Moreover, WER presupposes a "correct" transcription, which necessarily reflects a standardized orthography and a prestige language norm. For languages and dialects without a standardized written form, or for speakers who code-switch between multiple varieties, the concept of "ground truth" transcription is itself contested. When ASR systems are evaluated against reference transcriptions produced by speakers of standard varieties, the evaluation metric becomes a mechanism for enforcing linguistic conformity as a measurement instrument that encodes the values it purports to neutrally assess [25].

4.3. Language Model Priors as Policy

Language models embedded in speech AI systems encode assumptions about what constitutes "probable" speech. These priors are often learned from text corpora that share the same representational biases as speech training data, then assign higher probability to utterances that conform to dominant language patterns. A Swahili speaker in Nairobi who code-switches between Swahili and English receives lower confidence scores not because their speech is less fluent but because the model assigns lower probability to mixed-language sequences. An Akan speaker whose speech includes tonal distinctions absent from the model's phonemic inventory is systematically "corrected" toward the nearest English approximation.

These three levels of linguistic policy: data curation, metric selection, and model priors, interact to produce a system that does not merely fail to recognize certain voices; it *actively produces* certain voices as unrecognizable. This is the de-colonial insight: the exclusion is not the absence of inclusion but the presence of a normative order that constitutes some speech as unintelligible by design.

5. The Three Harms: Misrecognition, Misalignment, and Mistrust

Conventional evaluation of speech AI focuses primarily on transcription accuracy (e.g., WER), but this captures only one dimension of harm and often mis-specifies what counts as failure in multilingual, culturally situated use. In broader NLP, representational and allocational harms have been used to describe how systems reinforce stereotypes or distribute resources unevenly [26, 27]. The 3M taxonomy translates these insights into speech-specific, cross-cultural failure modes that are grounded in how people are heard, understood, and trusted in real interaction.

Drawing on the theoretical foundations outlined above, we introduce a taxonomy of three interrelated harms, the **3M framework**, which captures the full spectrum of ways speech AI systems fail Global South language communities.

Table 2 summarizes the 3M taxonomy as a diagnostic instrument: each failure mode implies distinct evaluation protocols and distinct design interventions.

Table 2: *The 3M taxonomy operationalized: definitions, how to detect, and why it matters.*

Failure mode	**What it is**	**How to detect / measure**	**Downstream harm**
Misrecognition	The system fails to produce an accurate transcript or intent (e.g., deletions/substitutions, refusals).	WER/CER; refusal rate; intent error; subgroup error gaps; acoustic condition sensitivity.	Denied access, delayed service, exclusion from voice-mediated infrastructure (health, safety, education).
Misalignment	The system produces plausible output but *misinterprets meaning* in culturally situated speech (idioms, honorifics, indirectness, code-switching).	Pragmatic adequacy ratings; culturally grounded eval sets; meaning-vs-wording annotation; implicature checks.	Loss of dignity, safety risks, harmful "corrections" toward prestige norms; cultural erasure through normalization.
Mistrust	Relational harm: users disengage because the system feels unreliable, disrespectful, or surveillant.	Abandonment/turn-taking breakdown; trust/agency scales; qualitative interviews; complaint logs; opt-out rates.	Feedback loop: reduced usage reduces repair signals, widening performance gaps and discouraging adoption.

5.1. Misrecognition

Misrecognition encompasses the most visible failures: incorrect transcriptions, dropped words, wrong-speaker attributions, and complete non-recognition (where the system produces no output at all). While WER captures some of these failures, misrecognition extends beyond word-level errors to include systematic patterns of erasure. Koenecke et al. found that five major ASR systems exhibited average WER differences of approximately 16 percentage points between African American and White speakers [3]. Martin and Tang document how these disparities persist and compound for speakers of African American Language [28].

For Global South language communities, misrecognition operates at an even more fundamental level. Speakers of Wolof, Igbo, or Haitian Creole encounter systems that do not merely misrecognize their speech; they encounter systems that deny the existence of their language as a supported category. This categorical exclusion, which is often the system's refusal to even *attempt* recognition, is a distinct form of misrecognition that WER cannot measure because there is no output to evaluate.

5.2. Misalignment

Misalignment occurs when a system produces a plausible transcription or response but *misinterprets meaning* in culturally situated speech. The words may be "right" while the intended force, social positioning, or referent is wrong; this makes misalignment invisible to WER and other surface-form metrics. Because it reflects mismatches in pragmatics and cultural context, misalignment is a key cross-cultural harm in speech AI

[29].

Three illustrative cases make this concrete. First, *idiomatic misalignment*: a Twi speaker uses "*wo maame awō wo*" (literally "your mother gave birth to you") as admiration, but the system renders it as a literal or offensive statement. Second, *honorific misalignment*: a Hindi speaker contrasts "*aap*" (formal you) with "*tum*" (informal you) to signal respect; collapsing these forms removes socially consequential meaning. Third, *code-switching misalignment*: multilingual speakers routinely mix languages within a single utterance [30]; systems that force a single-language interpretation erase the communicative strategy and can change intent.

Detecting misalignment requires evaluation methods that go beyond transcription accuracy. We propose three complementary approaches: (i) *pragmatic adequacy* ratings, where community evaluators judge whether the system captured intended meaning and social stance; (ii) *culturally grounded evaluation sets* that include idioms, honorifics, register shifts, and locally salient frames; and (iii) *implicature/intent checks* that test whether the system preserves implied meaning rather than only literal wording. These methods make misalignment measurable without reducing it to WER.

5.3. Mistrust

Mistrust is the experiential and relational harm that accrues when communities perceive, albeit correctly, that a speech AI system was not designed for them. It manifests in interactional signals such as abandonment, repeated corrections, opt-out behavior, and the choice to route around voice interfaces entirely. Mengesha et al. show that speakers of stigmatized English dialects often frame assistants as "not very intelligent," a coping strategy that nonetheless signals erosion of trust [31].

Mistrust creates an adoption and repair loop that worsens inequity. As users disengage, systems receive fewer error corrections and less representative interaction data; this reduces performance for the very groups already underserved, further deepening mistrust. In practice, mistrust is therefore not only a perception problem but a structural mechanism that entrenches initial exclusion.

Mistrust also has a surveillance edge case. For communities with histories of state or corporate monitoring, speech systems can be perceived as tools of extraction or control, even when they function accurately. In such contexts, "working" systems can still be rejected on safety and dignity grounds, especially if always-on listening, data retention, or unclear consent processes are present [32]. Designing for cultural competence therefore requires recognizing mistrust as a rational response to surveillance risk, not a defect in user attitudes.

The 3M taxonomy is designed to be *additive*, not substitutive: it complements WER and other quantitative metrics rather than replacing them. Its value lies in making visible the dimensions of harm that technical metrics alone cannot capture, and in grounding evaluation in the experiences of the communities affected by speech AI's linguistic policies.

6. Community-Led Responses and Existing Efforts

The theoretical critique advanced above is not merely academic; it is animated by existing community-driven initiatives that are already building alternatives to extractive speech AI development. These initiatives demonstrate that Global South language communities are not passive recipients of technological inclusion but active co-designers of speech systems grounded in their own linguistic and cultural contexts.

Masakhane, a pan-African grassroots research collective, has organized hundreds of researchers across the continent to develop NLP tools for African languages through participatory methods [33]. Rather than waiting for major technology companies to address African language representation, Masakhane members build datasets, train models, and establish evaluation standards rooted in community needs. The initiative embodies a *reciprocal* model of development in which communities retain ownership over their linguistic data and shape the research agenda.

UGSpeechData provides approximately 5,000 hours of validated speech across five Ghanaian languages, Akan, Ewe, Dagbani, Dagaare, and Ikposo [34]. The corpus was assembled through controlled crowdsourcing (a curated data-collection workflow with screening and expert validation), with distinct roles for speakers, validators, and transcribers [34]. Indigenous speakers contributed spontaneous image descriptions through a purpose-built Android application; recordings were screened and then evaluated by proficient speakers and language experts using multi-rater checks with defined conflict-resolution steps [34]. To support automatic speech recognition (ASR, the task of converting speech audio into text) and related analyses, a subset of validated audio was selected for transcription, reported as roughly 10% per language, which corresponds to about 100 hours per language, alongside language-appropriate input tooling, including customised keyboard support for orthographies that are not well served by standard layouts [34]. This example may suggest that resource constraints in Global South speech AI are not only a matter of volume; the more persistent limitation often appears to be the availability of governance processes that sustain quality and community legitimacy as datasets grow [34]. It also offers a practical pathway for aligning dataset construction with local language practices, whereas many widely used benchmarks remain shaped by assumptions that fit dominant-language norms more readily than multilingual community use.

Beyond dataset construction, **GhanaNLP has also operationalized Ghanaian language technologies in public-facing systems** that foreground everyday utility. For example, the *Khaya* application supports translation and (in recent iterations) speech capabilities for Ghanaian and other African languages, providing an accessible interface layer through which speakers can interact with models in culturally familiar terms [35, 36]. GhanaNLP also curates public repositories of language resources and models for Ghanaian languages (e.g., Akan, Dagbani, Ga), lowering the barrier for local experimentation and educational adoption [37]. Importantly, these initiatives reinforce our claim that "inclusion" is not only a matter of adding languages to corporate roadmaps, but of building *locally legible* language infrastructure: interfaces, governance practices, and maintenance pathways that allow communities to shape how speech AI is accessed, evaluated, and iteratively improved.

AI4Bharat has developed open-source language technology infrastructure for Indian languages, including the IndicNLPSuite of corpora, benchmarks, and pretrained models [38]. This initiative addresses the digital language divide for South Asian languages by building public infrastructure that any researcher or developer can use, avoiding the proprietary enclosures that characterize much of the speech AI industry.

Mozilla Common Voice provides an open platform for community-contributed speech data across more than 100 languages [39]. Its contribution model, where speakers voluntarily record and validate utterances, offers one pathway to address data scarcity, although scholars have noted the need for stronger governance mechanisms to ensure that community data are not extracted for commercial purposes without reciprocal benefit [40].

In Kenya, researchers at Microsoft Research Africa are advancing work on Kenyan and Swahili tribal languages, building culturally grounded language technologies that prioritize local knowledge systems. **Lesan.ai** is developing machine translation for Ethiopian languages, addressing one of the most linguistically diverse regions in the world.

Workshops, institutes, and dataset challenges as ecosystem infrastructure. Community capacity-building also occurs through venues that make African language work visible and citable. The *AfricaNLP* workshop series explicitly frames "unlocking local languages" as a research and community agenda, helping consolidate shared tasks, norms, and networks for African NLP [41]. Complementing workshops, practitioner-led institutes such as the Niger-Volta Language Technologies Institute develop speech and language resources for West African languages (e.g., ASR, language identification, and datasets for Yorùbá) through open repositories and community partnerships [42, 43]. Dataset challenges, including the AI4D African Language Dataset Challenge, similarly mobilize distributed participation to surface and improve underrepresented language data [44]. Finally, regional research networks (e.g., EthioNLP) support sustained scholarly visibility for Ethiopian languages, where linguistic diversity is substantial and the resource gap remains acute [45]. These ecosystem layers matter because they shift the "linguistic policy" of what counts as legitimate research output, who is recognized as an expert, and which language varieties receive iterative investment.

Indigenous languages in Latin America and the Americas. There are on-going community efforts in Latin America foreground the same structural problem: linguistic "low-resource" status is historically produced and institutionally reinforced. The *AmericasNLP* workshop series has helped organize shared tasks and benchmarks for Indigenous languages of the Americas, often in translation settings where Spanish acts as a high-resource pivot language [46]. Recent survey work also documents uneven NLP progress across Indigenous Latin American languages, emphasizing that the limiting factor is not only data volume but also institutional support, evaluation norms, and long-term stewardship practices [47]. This comparative lens reinforces our argument that cultural competence cannot be solved by a single multilingual model; it requires shifting the socio-technical conditions under which language varieties become visible, supported, and governable.

Afro-diaspora English varieties as a "within-English" resource divide. A useful comparative lens comes from Afro-diaspora English varieties across the Americas and the Caribbean, where colonial English has been taken up, transformed, and sustained through ethno-linguistic community practice—including African American English (AAE)[48], Jamaican Patwa and Caribbean Creole [49, 50], Barbadian English, and Belizean English [17]. Although English is typically treated as a "Winner" language in NLP resource taxonomies, these varieties often remain underrepresented in training data, evaluation sets, and product-facing speech interfaces that implicitly privilege standardized prestige norms. As a result, speakers can experience a familiar pattern of exclusion [3]: *misrecognition* (accent/dialect-specific error), *misalignment* (pragmatic and semantic misunderstanding of culturally situated expressions), and *mistrust* (disengagement after repeated failures). This illustrates a broader point central to our argument: cultural competence cannot be inferred from language-level coverage alone. Even when a language appears "well-resourced," colonial linguistic ideology can reproduce scarcity and marginalization at the level of dialect, creole, and community variation [51].

These initiatives share several features that distinguish them from conventional "inclusion" efforts: they are community-led rather than corporate-led; they prioritize community governance over data extraction; they define research agendas from local needs rather than external benchmarks; and they treat speakers as experts on their own languages rather than as data sources. The framework we propose in the next section formalizes these principles into a replicable design methodology.

7. A Participatory Framework for Culturally Competent Speech AI

Building on the theoretical analysis in Sections 2–3, the harm taxonomy in Section 4, and the community-led models in Section 5, we propose a participatory framework for designing and evaluating culturally competent speech AI. The framework comprises four interconnected pillars, each addressing a different phase of the speech AI lifecycle.

Recent work in culturally competent and culture-sensitive AI emphasizes that failures often arise not from isolated technical errors, but from *cultural incongruencies*: mismatches between a system's implicit cultural assumptions and the situated norms, values, and communicative practices of the communities it serves [29]. Complementing this, culture-sensitive and participatory approaches in public-sector AI stress that responsiveness requires more than representational inclusion; it requires iterative engagement with stakeholders, contextual accountability, and mechanisms for adjustment when harms surface in practice [52].

Our framework differs from these prior approaches in two key ways. First, rather than treating cultural competence as a high-level design aspiration or domain-specific requirement, we *operationalize* it through the 3M taxonomy as a concrete, recurring audit criterion: systems must be assessed not only for misrecognition, but also for misalignment (culturally situated meaning failures) and mistrust (relational and legitimacy failures). Second, we model the speech AI lifecycle as a *loop of repair and redress* rather than a one-time alignment exercise: participatory auditing informs community co-design choices, which shapes equitable deployment constraints, which then generates feedback signals that trigger further auditing and remediation. This loop makes cultural competence measurable, revisable, and governable over time, particularly in contexts where colonial linguistic hierarchies have historically shaped what speech is treated as "standard," "supported," and "intelligible." In speech settings, these incongruencies can surface even when transcription appears correct, because pragmatic meaning (e.g., honorifics, indirectness, idioms, code-switching) is culturally indexed and easily normalized away by dominant-language priors.

Figure 2 depicts the framework as a loop rather than a pipeline, emphasizing that cultural competence requires continuous repair. The 3M taxonomy functions as the auditing instrument that links evaluation to redesign and redress.

Participation is not inherently emancipatory. Without explicit safeguards, participatory processes can produce representational capture (where a narrow set of voices stands in for the whole), community-washing (symbolic involvement with-

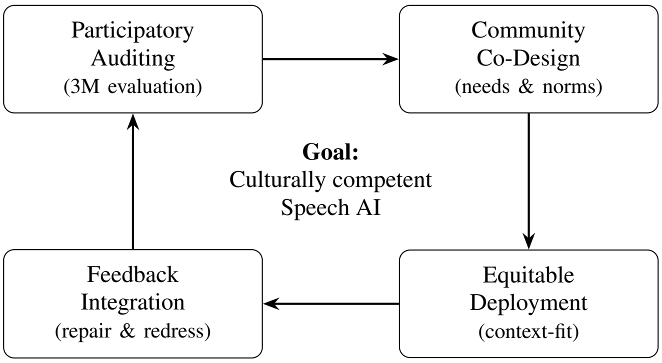

Figure 2: *A participatory framework for culturally competent speech AI. The four pillars form a continuous loop; participatory auditing uses the 3M taxonomy to surface misrecognition, misalignment, and mistrust and to guide repair.*

out authority), and extractive data collection that does not include withdrawal rights or benefit-sharing governance [40]. The framework therefore treats participation as a governance commitment rather than a methodological add-on.

7.1. Pillar 1: Participatory Auditing

Participatory auditing shifts evaluation authority from system developers to the communities whose voices the system is intended to serve. Rather than relying solely on benchmark datasets and quantitative metrics, participatory auditing involves community members in identifying, characterizing, and prioritizing system failures.

Concretely, participatory auditing proceeds as follows. First, community members—not external annotators—generate culturally situated test utterances that reflect real-world speech contexts, including code-switching, idiomatic expressions, honorifics, and domain-specific vocabulary (e.g., healthcare terminology in local languages). Second, community members evaluate system outputs against all three dimensions of the 3M taxonomy: transcription accuracy (Misrecognition), meaning fidelity (Misalignment), and subjective trust (Mistrust). Third, communities prioritize which failures are most consequential *in their context*, a distinction that external evaluators cannot make without situated knowledge.

This approach builds on the design justice principle that those most affected by a system's outcomes should have the greatest say in its evaluation [53], and addresses critiques that participatory methods must go beyond tokenistic inclusion to redistribute evaluative authority [40, 54].

7.2. Pillar 2: Community Co-Design

Community co-design involves language communities in shaping the data collection, annotation, and model development processes, not as consultants after the fact, but as partners from the outset. Drawing on Harrington et al.'s deconstructed community-based design methodology [55], co-design in this framework requires three commitments: (1) language communities define what counts as "accurate" and "appropriate" system behavior for their contexts; (2) annotation guidelines are developed collaboratively, reflecting local linguistic norms rather than external standards; (3) communities retain governance rights over the speech data they contribute, including the right to withdraw data and to shape how it is used.

The co-design process also attends to what we call *productive non-recognition*: cases in which communities may prefer that a system *not* recognize or record certain speech, for reasons of cultural protocol, privacy, or resistance to surveillance. This is a dimension of cultural competence that inclusion-focused frameworks typically overlook: true competence includes knowing when not to listen.

7.3. Pillar 3: Equitable Deployment

Equitable deployment addresses the governance of speech AI systems once they are built. It asks: who controls the system? Who benefits from its deployment? Who bears the costs of its failures? This pillar draws on Sambasivan et al.'s analysis of how fairness must be re-imagined across cultural contexts [56] and Mohamed et al.'s de-colonial AI framework [12].

Equitable deployment requires that communities have a role in deciding (a) whether a speech AI system is deployed in their context at all, (b) under what conditions it operates, (c) how failures are reported and remediated, and (d) how benefits (e.g. economic benefits from data and model improvements), are distributed. This is not a technical requirement but a governance one, and it necessarily involves institutional, legal, and political mechanisms beyond the scope of system design alone.

7.4. Pillar 4: Feedback Integration

Feedback integration closes the loop between deployment and development. It ensures that community-identified failures, cultural misalignments, and trust deficits are systematically channeled back into the system improvement process. Concretely, this involves mechanisms for users to (a) flag when a system has misrecognized their speech, (b) provide corrected transcriptions in their own terms, (c) indicate when the system has misunderstood meaning or intent, and (d) report when they feel the system was not designed for them.

These four pillars are interdependent: participatory auditing generates the situated knowledge that informs co-design; co-design produces systems that are more amenable to equitable deployment; equitable deployment creates the institutional conditions for sustained feedback integration; and feedback integration improves the system in ways that participatory auditing can verify.

8. Discussion and Future Work

This paper offers a theoretical and conceptual contribution; it does not present experimental results or empirical data from new studies. This is a deliberate choice. The speech AI community's understandable emphasis on empirical benchmarking can obscure the need for the kind of foundational rethinking we propose here: before we can measure cultural competence, we must articulate what it means; before we can audit for harm, we must develop a taxonomy of harms that extends beyond transcription accuracy; before we can design participatory systems, we must theorize participation in ways that attend to colonial histories and power asymmetries.

That said, the framework we propose is designed for empirical operationalization. We envision three lines of future work. First, a **cross-linguistic bias audit** applying the 3M taxonomy to evaluate commercial and open-source ASR systems across 3–5 Global South language communities, producing both quantitative disparity measurements and qualitative harm assessments. Second, a **participatory co-design study** in partnership with community organizations in West Africa, East Africa, and/or South Asia, testing the framework's four pillars in practice and generating empirically grounded design requirements for culturally competent speech AI. Third, a **systematic linguistic policy analysis** mapping the implicit and explicit language policies of major speech AI platforms against colonial language hierarchies, producing actionable policy recommendations for developers and regulators.

We acknowledge several limitations. Our framework is intentionally broad; its application to specific language communities will require contextual adaptation. The four-pillar structure simplifies what are, in practice, deeply entangled processes. And participatory methods, as Sloane et al. remind us, are not inherently emancipatory: they must be designed with care to avoid reproducing the power asymmetries they aim to address [40].

We also argue for *non-recognition* as a legitimate design outcome. In some contexts, opacity-by-design and selective listening are necessary for privacy, safety, and cultural protocol. Communities should be able to determine when systems listen, what gets stored, and when speech is intentionally *not* recognized or retained, especially in settings where surveillance is a credible risk [32].

Refusal and selective intelligibility as design practice. Our emphasis on non-recognition is not a retreat from inclusion, but an argument for *selective intelligibility*: communities should be able to determine when, how, and for what purposes their voices become machine-readable. This aligns with long-standing work in privacy and resistance that treats *obfuscation* and intentional friction as legitimate responses to extractive data regimes [57]. It also resonates with critiques of classification systems that produce harm precisely by forcing people into legible categories (e.g., automatic gender recognition), where refusal can be both protective and politically clarifying [58]. In this spirit, we position refusal not merely as a post-hoc user option, but as a *designable governance commitment*, a set of defaults, interfaces, and institutional guardrails that restrict capture, limit downstream reuse, and preserve the right to opacity when recognition would amplify risk or reproduce colonial linguistic hierarchies [59].

If speech AI is to fulfill its promise of bringing people and communities closer across languages and cultures, it must reckon with the colonial legacies that have historically determined who gets to speak, who gets understood, and who gets to build the systems that listen.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proc. International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[2] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 2426–2430.

[3] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Tober, H. R. Reeves, J. R. Grieco, E. R. Cerf, G. R. Brammer, and S. Goel, "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[4] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 6282–6293.

[5] R. Phillipson, *Linguistic Imperialism*. Oxford: Oxford University Press, 1992.

[6] N. ugĩ wa Thiong'o, *Decolonising the Mind: The Politics of Language in African Literature*. London: James Currey, 1986.

[7] P. Bourdieu, *Language and Symbolic Power*. Cambridge, MA: Harvard University Press, 1991.

[8] J. Rosa and N. Flores, "Unsettling Race and Language: Toward a Raciolinguistic Perspective," *Language in Society*, vol. 46, no. 5, pp. 621–647, 2017.

[9] N. Flores and J. Rosa, "Undoing Appropriateness: Raciolinguistic Ideologies and Language Diversity in Education," *Harvard Educational Review*, vol. 85, no. 2, pp. 149–171, 2015.

[10] S. M. Ali, "A Brief Introduction to Decolonial Computing," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 22, no. 4, pp. 16–21, 2016.

[11] L. Irani, J. Vertesi, P. Dourish, K. Philip, and R. E. Grinter, "Postcolonial Computing: A Lens on Design and Development," in *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, 2010, pp. 1311–1320.

[12] S. Mohamed, M.-T. Png, and W. Isaac, "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence," *Philosophy & Technology*, vol. 33, no. 4, pp. 659–684, 2020.

[13] D. Paffey, *Language Ideologies and the Globalization of "Standard" Spanish*. London: Bloomsbury Academic, 2012.

[14] A. Soares da Silva, "Pluricentricity and variation in the portuguese language," in *Pluricentric Languages in the Americas*, R. Muhr and A. Soares da Silva, Eds. Berlin: Peter Lang, 2022.

[15] S. Bird, "Decolonising Speech and Language Technology," in *Proc. International Conference on Computational Linguistics (COLING)*, 2020, pp. 3504–3519.

[16] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 6282–6293.

[17] R. Lippi-Green, *English with an Accent: Language, Ideology, and Discrimination in the United States*, 2nd ed. London: Routledge, 2012.

[18] P. Dourish and S. D. Mainwaring, "Ubicomp's Colonial Impulse," in *Proc. ACM Conference on Ubiquitous Computing (UbiComp)*, 2012, pp. 133–142.

[19] F. Fanon, *Black Skin, White Masks*. New York: Grove Press, 1952.

[20] D. E. Blasi, A. Anastasopoulos, and G. Neubig, "Systematic Inequalities in Language Technology Performance across the World's Languages," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 5486–5505.

[21] N. Couldry and U. A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media*, vol. 20, no. 4, pp. 336–349, 2019.

[22] D. Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective," in *Women, Science, and Technology*. Routledge, 2013, pp. 455–472.

[23] L. Suchman, "Located accountabilities in technology production," *Scandinavian Journal of Information Systems*, vol. 14, no. 2, pp. 7–19, 2002.

[24] N. Markl, "Language Variation and Algorithmic Bias: Understanding Algorithmic Unfairness and the Role of Language," in *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[25] A. B. Wassink *et al.*, "Uneven Success: Automatic Speech Recognition and Ethnicity-Related Dialects," *Speech Communication*, vol. 140, pp. 50–70, 2022.

[26] S. L. Blodgett, Q. V. Liao, A. Olteanu, R. Mihalcea, M. Muller, M. K. Scheuerman, C. Tan, and Q. Yang, "Responsible language technologies: Foreseeing and mitigating harms," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–3.

[27] S. L. Blodgett, S. Barocas, H. Daume, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in nlp," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[28] J. L. Martin and K. Tang, "Bias in Automatic Speech Recognition: The Case of African American Language," *WIREs Computational Statistics*, vol. 16, no. 2, p. e1650, 2024.

[29] V. Prabhakaran, R. Qadri, and B. Hutchinson, "Cultural incongruencies in artificial intelligence," *arXiv preprint arXiv:2211.13069*, 2022.

[30] A. DiChristofano, H. Shuster, S. Chandra, and N. Patwari, "Global Voices, Local Biases: Socioeconomic Disparities in Speech Recognition across Languages," *arXiv preprint arXiv:2305.11080*, 2024.

[31] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""I Don't Think These Devices Are Very Intelligent" – Examining Perceptions of Voice Assistants among Speakers of Stigmatized Dialects of English," in *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, 2021, pp. 1–13.

[32] S. Browne, *Dark matters: On the surveillance of blackness*. Duke University Press, 2015.

[33] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. H. Muhammad, S. K. Kabenamualu, S. Osei, F. Sackey *et al.*, "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2144–2160.

[34] I. Wiafe, J.-D. Abdulai, A. O. Ekpezu, R. D. Helegah, E. D. Atsakpo, C. Nutrokpor, F. B. P. Winful, and K. K. Solaga, "Advancing automatic speech recognition for low-resource ghanaian languages: Audio datasets for akan, ewe, dagbani, dagaare, and ikposo," *Data in Brief*, vol. 61, p. 111880, 2025.

[35] "Khaya: African language translation & speech (app / documentation)," https://translate.ghananlp.org/, accessed 2026-02-21.

[36] "Ghananlp," https://ghananlp.github.io/, accessed 2026-02-21.

[37] "Ghananlp: Ghanaian nlp datasets & models (repository)," https://github.com/GhanaNLP/ghanaian-nlp-datasets-models, accessed 2026-02-21.

[38] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961.

[39] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proc. Language Resources and Evaluation Conference (LREC)*, 2020, pp. 4218–4222.

[40] M. Sloane, E. Moss, O. Awomolo, and L. Forlano, "Participation Is Not a Design Fix for Machine Learning," in *Proc. Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022, pp. 1–6.

[41] "Africanlp 2020: Unlocking local languages (workshop cfp)," https://easychair.org/cfp/africanlp2020workshop, accessed 2026-02-21.

[42] "Niger-volta language technologies institute (organization)," https://github.com/niger-volta-lti, accessed 2026-02-21.

[43] "Niger-volta-lti: Yorùbá voice dataset (repository)," https://github.com/Niger-Volta-LTI/yoruba-voice, accessed 2026-02-21.

[44] "Ai4d african language dataset challenge," https://zindi.africa/competitions/ai4d-african-language-dataset-challenge, accessed 2026-02-21.

[45] "Ethionlp: Ethiopian natural language processing research community," https://ethionlp.github.io/, accessed 2026-02-21.

[46] "Americasnlp: Workshop on nlp for indigenous languages of the americas," https://turing.iimas.unam.mx/americasnlp/, accessed 2026-02-21.

[47] , "Nlp progress in indigenous latin american languages," *arXiv*, 2024, https://arxiv.org/abs/2404.05365.

[48] J. R. Rickford, *African American Vernacular English: Features, Evolution, Educational Implications*. Oxford, UK: Blackwell, 1999.

[49] P. L. Patrick, *Urban Jamaican Creole: Variation in the Mesolect*. Amsterdam: John Benjamins, 1999.

[50] J. A. Holm, *Pidgins and Creoles, Volume 1: Theory and Structure*. Cambridge, UK: Cambridge University Press, 1988.

[51] R. Lippi-Green, *English with an Accent: Language, Ideology, and Discrimination in the United States*, 2nd ed. New York, NY: Routledge, 2012.

[52] P. Ahrweiler, "Towards culture-sensitive, responsive, and participatory ai," in *Participatory Artificial Intelligence in Public Social Services: From Bias to Fairness in Assessing Beneficiaries*. Springer, 2025, pp. 277–306.

[53] S. Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press, 2020.

[54] A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed, "Power to the People? Opportunities and Challenges for Participatory AI," in *Proc. Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022, pp. 1–6.

[55] C. Harrington, S. Erete, and A. M. Piper, "Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements," in *Proc. ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2019, pp. 1–25.

[56] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining Algorithmic Fairness in India and Beyond," in *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 315–328.

[57] F. Brunton and H. Nissenbaum, *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press, 2015.

[58] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–22, 2018.

[59] J. Cunningham, G. Benabdallah, D. Rosner, and A. Taylor, "On the grounds of solutionism: Ontologies of blackness and hci," *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 2, pp. 1–17, 2023.