

Sovereignty Theater: Why Sovereign AI Falls Short of Cultural Self-Determination

Bruria Friedman Feldman*, Tyler Smith*, Victor Y. Wu*, Yuanhao Zou*

Stanford University

bruriaff@stanford.edu, ts95352@stanford.edu, victorywu@stanford.edu, zouyh@stanford.edu

Abstract

Sovereign AI has emerged as a major policy movement, with nearly 130 government-backed initiatives across more than 50 countries justified in part by the promise of cultural and linguistic self-determination. States seek national AI systems that reflect local languages, values, and knowledge traditions rather than importing the assumptions of foreign technology providers. Yet sovereign AI initiatives overwhelmingly depend on hardware, base architectures, and training methodologies originating from a concentrated set of U.S. vendors, and recent empirical research suggests that the dominant technical approach—fine-tuning a foreign-developed base model on local data—does not reliably produce culturally aligned outputs. We argue that sovereign AI in its current commercial form merely produces *sovereignty theater*: operational sovereignty (control over compute location and jurisdictional access) without epistemic sovereignty (genuine reflection of local cultural values and knowledge traditions). Bridging the sovereign AI policy literature with empirical research on LLM cultural alignment, and drawing on case analyses of ALLaM (Saudi Arabia) and SEA-LION (Singapore/Southeast Asia), we propose an analytical framework distinguishing infrastructural sovereignty, epistemic sovereignty, and participatory legitimacy. Importantly, our framework formalizes a bottleneck condition in which overall sovereignty is constrained by its weakest dimension. Genuine sovereign AI requires reorienting the project from a commercial product to a participatory endeavor that includes citizens, linguistic minorities, and civil society in the value-laden choices that shape model behavior. We close with a lifecycle decision matrix for policymakers and procurement teams to diagnose sovereignty theater in practice. Although developed in the context of sovereign AI, sovereignty theater offers a portable analytical lens for examining claims to technological sovereignty more broadly.

1 Introduction

Governments around the world are increasingly claiming that national AI systems must reflect their own languages, cultures, and values rather than importing the assumptions of foreign technology providers. South Korea’s NAVER Cloud defines sovereign AI as AI that “deeply understands a nation’s or region’s culture, history, and values by leveraging

their own data and infrastructure” (Dell Technologies 2025). India describes its flagship platform, Sarvam AI, as “building AI systems rooted in local languages, datasets, and regulatory frameworks” to ensure that “innovation aligns with national priorities and societal needs” (Press Information Bureau 2026). These are not isolated cases. By early 2026, nearly 130 government-backed sovereign AI projects had been launched across more than 50 countries (Chavez 2026), and McKinsey projects the sovereign AI market could reach \$600 billion by 2030 (Ustun et al. 2025). National strategies invoke a heterogeneous mix of motivations—national security, economic competitiveness, and more—among which cultural and linguistic self-determination is a particularly important and politically resonant strand (Varela Sandoval et al. 2026; Chavez 2026).

Yet there is a paradox at the heart of this movement. Even initiatives that pursue domestic model development overwhelmingly depend on hardware, base architectures, and training methodologies originating from a small number of U.S. technology firms. Many countries go further, acquiring sovereign AI capabilities through turnkey commercial packages sold by these same vendors. At the same time, a growing body of empirical research on cultural alignment in large language models (LLMs) suggests that the core technical approach underlying these packages—fine-tuning a foreign-developed base model on local data—does not reliably produce culturally aligned outputs (Agarwal et al. 2025). Together, these two facts motivate the central question of this paper: does sovereign AI, as currently produced and procured, deliver on its cultural promise?

We argue that it does not. Sovereign AI, in its current commercial form, primarily produces what we term *sovereignty theater*: it delivers *operational sovereignty*—control over where computation occurs, where data resides, and which jurisdiction governs access—without securing *epistemic sovereignty*—ensuring that the system’s outputs genuinely reflect the cultural values, moral frameworks, and knowledge traditions of the population it is intended to serve. We trace this gap to two interconnected drivers. The first is technical: pretraining data composition, model architecture, tokenization, and alignment processes embed upstream epistemic assumptions that downstream fine-tuning often does not dislodge. The second is institutional: the procurement structures through which sovereign AI is acquired

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

channel decision-making through state-vendor relationships, excluding citizens, linguistic minorities, and civil society from the value-laden choices that shape model behavior.

This paper's contribution is to bridge two literatures that have developed largely in parallel: scholarship on sovereign AI and digital sovereignty as matters of policy and geopolitics, and empirical research on cultural alignment and misalignment in LLMs. We propose an analytical framework distinguishing infrastructural sovereignty, epistemic sovereignty, and participatory legitimacy, and apply it to the commercial sovereign AI ecosystem. We then draw on empirical evidence from the cultural alignment literature and on case studies of two sovereign AI initiatives, ALLaM and SEA-LION, to ground the analysis. We argue that achieving sovereignty requires overcoming a bottleneck condition: infrastructural investment cannot compensate for epistemic or democratic deficits. Finally, we operationalize this framework as a lifecycle decision matrix that policymakers and procurement teams can use to diagnose sovereignty theater across the procurement, pre-development, development, and deployment stages of sovereign AI initiatives. Although we develop these concepts in the context of sovereign AI, we intend sovereignty theater as a reusable analytical lens: a way of asking, of any claim to technological sovereignty, in what sense sovereignty is being claimed, by whom it is enacted, and how it is legitimized.

2 Background: Sovereign AI as Policy and Product

2.1 The Policy Landscape

Since the late 2010s, scholars have discussed the notion of digital sovereignty: a state's capacity to exercise control over its digital infrastructure, data flows, and technological systems (Couture and Toupin 2019). The concept initially centered on data localization and platform governance and became prominent in European security discourse as governments confronted dependence on foreign cloud providers and digital platforms (Bellanova, Carrapico, and Duez 2022). The emergence of LLMs accelerated the shift toward discussions about "AI sovereignty." NVIDIA CEO Jensen Huang popularized the term in 2023 when he argued that every nation needs the capability to produce its own intelligence (Dale 2025). As of 2026, nearly 130 government-backed sovereign AI projects are active across more than 50 countries, a threefold increase from roughly 40 projects in 30 countries just two years earlier (Chavez 2026).

The term "digital sovereignty" is contested: it is invoked both by democratic states seeking to reduce foreign technology dependence and by authoritarian regimes—China's "cyber sovereignty," Russia's "sovereign internet," Iran's invocation of digital sovereignty—to justify content controls and surveillance (G'sell 2025). We engage sovereign AI in the first register, as a claim to collective cultural and epistemic self-determination, while recognizing that initiatives lacking participatory and contestability commitments may slide toward the authoritarian form. We return to this tension in Section 5.3.

In recent years, scholars have proposed conceptual frameworks for analyzing these developments. Singh and Sengupta (2025) propose a four-pillar framework defining sovereign AI along the dimensions of data, compute, models, and norms. They argue that sovereign AI is best modeled along a spectrum rather than as an on/off condition, with national positions varying according to fiscal capacity, institutional depth, and the strategic balance a country strikes between self-reliance and openness.

Language and culture are both central in the political justification of the current wave of sovereign AI. For example, as discussed earlier, India's Sarvam AI and South Korea's NAVER Cloud both frame their initiatives in explicitly cultural terms (Press Information Bureau 2026; Dell Technologies 2025). Meanwhile, the European Commission explicitly ties its multilingual AI services to "European sovereign AI" and is constructing an "EU institutional LLM" to support low-resource official languages such as Irish, Maltese, and Latvian (European Commission, Directorate-General for Translation 2025). Across these initiatives, cultural and linguistic self-determination figures prominently among the justifications offered for sovereign AI investment, even as other rationales—security, industrial strategy, and reduced dependency—are typically advanced in parallel (Varela Sandoval et al. 2026).

2.2 The Commercial Ecosystem

Despite this language of self-determination, sovereign AI initiatives overwhelmingly depend on hardware, base architectures, and training methodologies originating from a concentrated set of technology vendors mostly headquartered in the United States. The Center for a New American Security (CNAS) Sovereign AI Index finds that roughly 70 percent of tracked sovereign AI projects involve at least one foreign partner, with U.S. companies appearing in four-fifths of those partnerships (Chavez, Chilukuri, and Scanlon 2026). The geography of investment is equally lopsided, with the Middle East and East Asia together absorbing more than 80 percent of disclosed sovereign AI spending—and within that share, the United Arab Emirates and Japan alone responsible for roughly two-thirds of disclosed totals (Chavez, Chilukuri, and Scanlon 2026).

Moreover, a growing number of countries acquire sovereign AI capabilities through turnkey commercial packages: NVIDIA's "AI Factories," Microsoft's "Sovereign Cloud," Oracle's sovereign cloud regions, and HPE's "NVIDIA AI Computing" offerings each bundle GPU infrastructure, local hosting, and fine-tuning capabilities under domestic jurisdictional control (NVIDIA Corporation 2026a, 2025, 2026b; Phillips 2026; Oracle Corporation 2026; Hewlett Packard Enterprise 2025).

Consulting firms have moved to frame sovereign AI as a strategic market opportunity: McKinsey, for example, projects sovereign AI could represent a \$600 billion market by 2030, with up to 40 percent of AI workloads moving to sovereign environments (Ustun et al. 2025).

In practice, most commercial "sovereign AI" packages are built on foreign hardware (usually NVIDIA GPUs), a foreign base model or training toolkit, and some fine-tuning

infrastructure (Meinhardt et al. 2026). What “sovereign” really guarantees is thus where the computation happens and who has legal access to the data, but it does not guarantee what worldview the model carries. The architectures, training data, and alignment methods embedded in these systems still reflect the assumptions of whoever built them. There is an important difference between sovereign cloud (controlling where data lives) and sovereign AI (controlling how intelligence is built) (Ustun et al. 2025).

A parallel critique has emerged in the cloud services context. Sukumar (2025) argues that hyperscalers’ “sovereign cloud” offerings to European customers deliver cosmetic compliance with sovereignty rhetoric while leaving underlying dependencies on U.S. vendor infrastructure largely intact—a pattern he terms “sovereignty washing.” The AI-specific dynamic we examine here extends this concern to a deeper layer of the stack, raising the additional question of whether “sovereign” models meaningfully reflect local values and knowledge traditions.

This concern is sharpened by U.S. extraterritorial legal authorities. The CLOUD Act (2018) authorizes U.S. law enforcement to compel U.S.-headquartered providers to produce data they hold regardless of where it is physically stored, and Section 702 of the Foreign Intelligence Surveillance Act permits surveillance directed at non-U.S. persons whose communications transit U.S.-controlled infrastructure or services. Because the dominant commercial “sovereign AI” offerings are sold by U.S. firms or built atop U.S.-developed base models and cloud platforms, these jurisdictional hooks frequently remain even when compute and data are physically localized within the purchasing country (Sukumar 2025; Chavez 2026). Thus, even operational sovereignty is less commercially attainable than vendor marketing suggests.

3 Analytical Framework: Operational vs. Epistemic Sovereignty

3.1 The Epistemic Gap

The concept of sovereign AI is often framed in terms of control: for instance, where data is stored or where computation takes place. But as Roberts (2024) argues in the broader context of digital sovereignty, sovereignty is not merely a descriptive question of control, but a normative claim to legitimate authority. Indeed, in the context of AI, controlling where a model runs is not the same as controlling what the model represents or whose values it encodes. A system may be locally hosted and nationally regulated yet still reflect epistemic assumptions that originate elsewhere.

To capture this distinction, we build on the dual-axis framework introduced by de Ocariz Borde (2025) and distinguish between operational sovereignty and epistemic sovereignty. Operational sovereignty refers to control over the infrastructure of AI systems: data residency, compute location, access restrictions, and compliance with domestic legal regimes. This is the dimension most readily delivered by contemporary “sovereign AI” offerings. Governments can ensure that models are deployed within national borders, that data does not leave the jurisdiction, and that access is medi-

ated through domestic institutions. In this sense, operational sovereignty is the most legible dimension and the one most readily marketed by commercial vendors.

Epistemic sovereignty, by contrast, concerns the substantive content of model behavior: whether the system’s outputs reflect the cultural values, moral frameworks, and social norms of the population it is intended to serve. This is the dimension invoked by the political rhetoric of sovereign AI, with its promise of systems that “understand” national culture and priorities, but it is far less reliably achieved in practice. The key claim of this paper is that current sovereign AI models systematically deliver operational sovereignty without securing epistemic sovereignty.

The persistence of the epistemic gap is rooted in the technical structure of contemporary AI systems, where core assumptions are established early in the model development pipeline and are difficult to dislodge through downstream interventions. Several mechanisms are particularly important.

First, pretraining data and tokenization establish a Western-skewed substrate. Web-scale corpora are overwhelmingly English-language and Western-produced, encoding both linguistic patterns and assumptions about social norms and moral priorities. Tokenization schemes optimized for high-resource languages compound this by fragmenting low-resource or morphologically rich languages.

Second, alignment processes and base model priors entrench these defaults. Reinforcement learning from human feedback introduces normative filtering through annotators concentrated in specific geographic and cultural contexts, propagating particular interpretations of helpfulness and appropriateness throughout the model. Fine-tuning on local data—the standard sovereign AI intervention—operates on top of these deeply embedded representations, adjusting surface behavior, vocabulary, and style without dislodging the underlying reasoning patterns.

In sum, the dominant technical approach to sovereign AI—deploying a foreign-developed base model and fine-tuning it on local data—systematically falls short of epistemic sovereignty. Operational control can be localized relatively late in the pipeline, at the level of deployment and infrastructure. Epistemic content, however, is largely determined upstream, at stages that remain concentrated in a small number of global actors. This asymmetry produces what we term the *epistemic gap*: a structural divergence between the political promise of sovereign AI and the technical realities of how such systems are built.

3.2 The Democratic Deficit

If sovereignty, as Roberts (2024) argues, is ultimately a claim to legitimate authority, then many sovereign AI initiatives also suffer from a democratic deficit. Although these projects localize infrastructure and jurisdiction, they rarely democratize the normative decisions embedded in AI systems. Core choices—what models to procure, what data to use, which languages to prioritize, and what values to operationalize—are typically made through centralized partnerships between states, vendors, and technical elites, with limited participation from affected communities.

This exclusion matters because deliberative democratic theory has long argued that legitimacy depends on meaningful participation by those subject to collective decisions (Dryzek 2002; Gutmann and Thompson 2000; Benhabib 1994). In the context of AI systems that increasingly mediate education, public administration, healthcare, and communication, decisions about alignment, evaluation, and acceptable outputs become exercises in public value ordering. A sovereign AI system that claims to represent a population while excluding that population from these processes offers only a thin account of legitimacy.

The problem becomes especially important in multilingual and internally diverse societies. Appeals to “national values” ignore disagreements over language, religion, gender, and political identity. Scholarship on Indigenous data sovereignty shows that state-led governance can reproduce domination when communities lack authority over how their knowledge is collected and used (Kukutai and Taylor 2016; Mohamed, Png, and Isaac 2020). Likewise, Png (2022) argues that Global South actors are often formally included yet structurally marginalized in AI governance. Sovereign AI may therefore be sovereign for the state while remaining non-sovereign for subnational communities.

This creates the risk of elite capture. Procurement and governance decisions are frequently controlled by politically connected actors whose interests diverge from those of affected populations. Sovereign AI systems may privilege official languages, encode majoritarian norms, or prioritize state legibility over cultural nuance while presenting themselves as expressions of collective self-determination (Metcalfe 2026). Technically complex governance environments are particularly vulnerable to such concentration of authority (Taeihagh 2021).

At the same time, greater representational inclusion may generate privacy and surveillance risks, especially for small or politically vulnerable groups. Capturing minority languages and cultural practices often requires granular and sensitive community data, increasing the risk of reidentification or political misuse. As Nissenbaum (2004) argues through the framework of contextual integrity, privacy concerns depend not simply on secrecy, but on whether information flows accord with appropriate social norms.

Sovereign AI may therefore fail democratically in two opposite ways: by excluding marginalized communities from representation, or by incorporating them under conditions that threaten their autonomy and safety. For this reason, epistemic sovereignty requires more than domestic infrastructure or localized fine-tuning. It requires institutional mechanisms through which affected communities can meaningfully participate in determining how models are trained, aligned, evaluated, and governed. Recent work on impacted-stakeholder participation in AI governance is especially important because it treats participation as a response to structural marginalization and group-based harms (Kaminski and Malgieri 2025).

In sum, the democratic deficit helps explain why operational sovereignty so often fails to produce epistemic sovereignty. Sovereign AI may localize infrastructure, but unless it is also built as a participatory political project, it

risks reproducing the very hierarchies it claims to overcome.

4 Evidence of Epistemic Gaps

With the operational/epistemic distinction in hand, we now turn to empirical evidence on how sovereign AI initiatives actually perform along these dimensions.

4.1 The Problem that Sovereign AI Tries to Address

A growing body of empirical research demonstrates that LLMs exhibit systemic patterns of cultural misalignment (Buyl et al. 2024). Across five generations of GPT models, outputs cluster around the value profiles of English-speaking and Protestant European societies on the Inglehart–Welzel World Cultural Map, remaining distant from African-Islamic regions (Tao et al. 2024). The misalignment extends to linguistic expression: studies of GPT-4o mini and Gemma 7B in Yoruba and Malayalam find adequate surface translation but failure to capture cultural nuance, with models reverting to Western-aligned interpretations (Dawson et al. 2024). Prompting in different languages does not, on its own, elicit culturally specific value systems (Arora, Kaffee, and Augenstein 2022).

4.2 Regional Models Have Failed to Achieve Epistemic Sovereignty

Growing interest in sovereign and regional AI initiatives responds to this baseline misalignment. Yet existing evidence suggests that recent efforts have had limited success in achieving what may be called epistemic sovereignty: the capacity to represent and organize knowledge according to locally grounded interpretive frameworks rather than inherited external priors.

Drawing from psychological and anthropological literature, Pawar et al. (2025) define cultural awareness in LLMs as a model’s ability to understand how context varies across societies. Meanwhile, cultural alignment refers to configuring systems to reflect the norms and values of their intended communities (Masoud et al. 2023). Most sovereign AI initiatives pursue this goal through downstream localization strategies, particularly fine-tuning foundation models on local-language corpora and evaluating them using culturally specific benchmarks. These approaches, however, largely leave intact the upstream epistemic structures embedded during pretraining.

Studies of Indic language models illustrate this limitation clearly. Systems such as ARYABHATTA-8B, AIRAVATA-7B, and GAJENDRA-7B are fine-tuned on Indian-language and culturally relevant data atop foreign base models such as LLaMA-2. While these models improve performance on certain culturally specific tasks, they continue to reproduce value distributions closer to Western societies on dimensions such as self-expression and individualism. Moreover, improvements on implicit cultural questions remain marginal, with some models showing gains of only 1% relative to their base models (Agarwal et al. 2025). Fine-tuning therefore modifies surface-level behavior without fundamentally

transforming the underlying representational structures inherited from upstream systems. While the most systematic evidence on this point comes from a single evaluation of Indic models (Agarwal et al. 2025), it converges with the broader pattern documented in Section 4.1, in which neither prompting in local languages (Arora, Kaffee, and Augenstein 2022) nor surface-level localization (Dawson et al. 2024) dislodges Western-centric value distributions.

A second limitation concerns how cultural alignment itself is evaluated. Existing benchmarks rely heavily on methodologies grounded in standardized tasks, statistical similarity measures, and quantifiable outputs. Such approaches treat culture as a stable and measurable property reducible to benchmark performance. Yet interdisciplinary scholarship increasingly argues that cultural meaning is contested and context-dependent (Qadri et al. 2025). By reducing culture to what can be formalized and scored, benchmark-driven evaluation fails to capture the pluralistic and situated dimensions of social knowledge.

These two limitations explain why sovereign models fall short of their epistemic promises. The following case studies demonstrate how these limitations manifest institutionally within contemporary sovereign AI initiatives.

4.3 Case Studies: What Sovereignty Looks Like on the Ground

The cases of ALLaM and SEA-LION provide concrete illustrations of how sovereign AI is currently instantiated in practice. Emerging from distinct political and regional contexts—the former from Saudi Arabia and the latter from Singapore/Southeast Asia—both initiatives are framed as sovereign efforts to represent languages and cultures in contemporary AI systems. At the level of rhetoric, each invokes a promise of sovereignty in both its operational and epistemic senses: systems that not only operate within national or regional jurisdictions, but more importantly, also reflect the linguistic, cultural, and normative particularities of the populations they are meant to serve. Yet, when examined more closely, both cases reveal a gap between this promise and its realization, especially on the epistemic level.

The Saudi initiative, ALLaM, is presented as a flagship effort to advance Arabic-language technology and to support the cultural values of the Arabic-speaking world. Its stated ambition extends beyond improving linguistic performance to encompass the representation of culturally appropriate norms, social practices, and value systems. In this framing, sovereignty is articulated in civilizational terms: Arabic language and culture are positioned as domains requiring their own AI infrastructure, capable of counterbalancing the dominance of English-centric models (Alrajjal 2025). However, the system that is actually built reflects a more hybrid configuration. ALLaM relies on a standard transformer-based architecture and a two-stage training pipeline in which large-scale English pretraining precedes subsequent training on mixed Arabic and English data (Humain AI 2024). While this approach enhances performance and efficiency, it also embeds the model within an epistemic pipeline that originates outside the cultural context it is meant to represent. The model’s architecture, training strategy, and underlying

toolchain are inherited from global AI ecosystems, particularly those centered in the United States. As a result, the model may improve performance on Arabic benchmarks without fully overcoming upstream representational priors.

This pattern is reinforced by the limited role of public participation in the system’s development and evaluation. Cultural alignment is operationalized through narrowly scoped processes, such as small annotator pools and synthetic dataset generation, rather than through broad-based or participatory mechanisms. Benchmarks used to evaluate ALLaM reduce cultural knowledge to simplified, standardized formats, relying on minimal human validation and offering little room for disagreement, contestation, or contextual interpretation. The “cultural values” the system claims to encode are thus not collectively articulated, but inferred and stabilized through centralized, technocratic procedures. ALLaM exhibits high operational sovereignty but low epistemic sovereignty: its claim to represent cultural values is not matched by participatory or pluralistic processes of validation.

A parallel, though somewhat more pluralistic, configuration can be observed in the Southeast Asian case of SEA-LION (Ng et al. 2025). Developed as part of Singapore’s national AI strategy, SEA-LION is framed as a regional response to the linguistic and cultural diversity of Southeast Asia. Its stated goals emphasize support for low-resource languages, sensitivity to regional contexts, and the development of inclusive AI systems capable of operating across multilingual environments. Unlike ALLaM’s civilizational framing, SEA-LION foregrounds diversity as its central organizing principle, positioning itself as a model for a heterogeneous linguistic landscape rather than a singular cultural domain.

In practice, SEA-LION shows similar limitations in epistemic sovereignty: its developers adapt existing models and datasets in high-resource languages, most often English. In the pretraining stage, SEA-LION was built through continued pretraining and fine-tuning on Llama-3.1-8B-Instruct and Gemma-2-9B—foundation models developed by Meta and Google, respectively—and Southeast Asian language data (Ng et al. 2025). This approach enables rapid development and strong multilingual performance, but it also inevitably entails inheriting the epistemic assumptions, architectures, and training regimes of upstream systems. As in the ALLaM case, the result is a system that is regionally specialized, but epistemically compromised. Linguistic coverage and domain knowledge are expanded through the fine-tuning process, but the underlying structure of knowledge representation remains dependent on architectures and priors outside the sovereign domain of the Southeast Asian community.

The evaluation framework associated with SEA-LION, SEA-HELM, reflects a more advanced attempt to address these limitations (Susanto et al. 2025). It incorporates multiple languages, diverse task types, and datasets designed to capture cultural and contextual variation. It also involves collaboration with native speakers and regional contributors, suggesting a greater degree of local engagement than in the ALLaM case. Nevertheless, this engagement remains limited in scope, authority, and epistemic “thickness.” Native

speakers are involved as annotators or dataset contributors, rather than as participants in the governance of evaluation itself. They do not determine the structure of tasks, the criteria for correctness, or the resolution of conflicting interpretations.

Participation is thus consultative rather than deliberative, and epistemic authority remains centralized within technical and research institutions. Read against the framework proposed by Qadri et al. (2025), this limitation becomes especially pronounced. Engaging with Clifford Geertz’s methodological contribution to symbolic anthropology, Qadri et al. (2025) argue that meaningful cultural representation requires “thick” evaluation grounded in situated interpretation, plurality of perspectives, and community-defined criteria of appropriateness. Seen in this light, SEA-HELM, despite its expanded scope, remains epistemically thin as it operationalizes culture through standardized tasks and pre-specified evaluation schemas, thereby reducing cultural knowledge to what can be formalized and scored. In doing so, it falls short of the “thick description” necessary for achieving epistemic sovereignty. Achieving this would require replacing limited participation in data production with genuine involvement in the interpretive and normative processes through which cultural meaning is constituted.

Both cases thus succeed on the most visible dimensions of AI inequality—language coverage and infrastructural control, qualifying as “AI polyglots” in de Ocariz Borde’s words (de Ocariz Borde 2025). But they struggle to transform the deeper structures through which knowledge is defined and governed. In both cases, cultural representation is operationalized through evaluative schemas that privilege measurability over interpretive richness and plurality.

5 Discussion and Synthesis

5.1 Modeling the Components of AI Sovereignty

The four-pillar framework proposed by Singh and Sengupta (2025) provides a useful starting point for analyzing sovereign AI as a problem of infrastructural capacity. However, as the preceding discussions demonstrate, it collapses epistemic and democratic dimensions into a thin notion of “normative alignment” and treats sovereignty as an additive function of commensurable inputs. To address these limitations, we propose a revised model that distinguishes three analytically distinct dimensions: infrastructural sovereignty, epistemic sovereignty, and participatory legitimacy. Infrastructural sovereignty in this formal framework corresponds to what Section 3.1 termed operational sovereignty; we adopt the label “infrastructural” here to mark its role as a quantifiable pillar aggregating data, compute, and model autonomy.

Overall structure. We define AI sovereignty as a function of these three dimensions:

$$S^{AI} = \mathcal{F}(I, E, P) \quad (1)$$

where I denotes infrastructural sovereignty, E epistemic sovereignty, and P participatory legitimacy.

Rather than treating these dimensions as fully substitutable, we model sovereignty as a bottleneck condition:

$$S^{AI} = \min\{I, E, P\} \quad (2)$$

This reflects the claim that deficits in epistemic or democratic legitimacy cannot be compensated by gains in infrastructure.

Infrastructural sovereignty. As Singh and Sengupta (2025) suggest, infrastructural sovereignty depends on data D , compute C , and model autonomy M , which we aggregate with weights ω_D , ω_C , and ω_M :

$$I = \omega_D D + \omega_C C + \omega_M M, \quad \sum_i \omega_i = 1 \quad (3)$$

Epistemic sovereignty. Building on the conceptual discussion above, we treat epistemic sovereignty as distinct from operational sovereignty. Whereas operational sovereignty concerns control over infrastructure, deployment, and jurisdiction, epistemic sovereignty concerns authority over the production, interpretation, and validation of the knowledge embedded in AI systems. As argued in the preceding section, contemporary sovereign AI initiatives often localize deployment while leaving upstream model priors, evaluative criteria, and representational assumptions substantially unchanged. Epistemic sovereignty therefore cannot be reduced to local hosting, fine-tuning on domestic data, or the symbolic inclusion of national languages. It depends on the institutional conditions under which meaning is authored and authorized. And like infrastructural sovereignty, epistemic sovereignty is best understood as a continuum rather than a binary.

To formalize this dimension, we define epistemic sovereignty over three core components: thickness of evaluation (T), epistemic authorship (A), and contestability (Q). These capture whether a system’s knowledge claims are locally meaningful, locally governed, and open to revision.

Drawing on Qadri et al. (2025), who engage Geertz’s methodological contribution to anthropology, we define thickness of evaluation (T) as the epistemic quality of evaluative methods. It measures whether cultural representation is assessed through contextual, interpretive, and socially situated procedures rather than thin proxies such as binary correctness, decontextualized benchmarks, or surface-level linguistic fluency. Drawing on traditions of “thick description” in anthropology and recent work on cultural evaluation in AI, this dimension emphasizes that meaning is often ambiguous, relational, and context-dependent. A system scores higher on T when its evaluation regime can register nuance, disagreement, and appropriateness across varied social settings rather than collapsing culture into fixed labels or multiple-choice responses.

Epistemic authorship (A) concerns who determines what counts as valid knowledge within the system. This includes authority over the selection of training corpora, benchmark design, taxonomies, alignment criteria, and the broader interpretive frames through which outputs are judged. The concept draws on science and technology studies, postcolonial critiques of knowledge production, and the literature on

epistemic injustice, all of which emphasize that knowledge is shaped by those who possess classificatory power (Mohamed, Png, and Isaac 2020). A system scores higher on A when affected communities, local institutions, or regionally grounded experts exercise meaningful control over these substantive standards rather than merely supplying data to externally structured pipelines.

Contestability (Q) measures whether outputs, labels, and evaluative standards can be challenged and revised. Because cultural meaning is dynamic and frequently disputed, epistemic sovereignty cannot rest on static benchmarks alone. It requires mechanisms through which affected groups can contest misrepresentation, appeal harmful outputs, and prompt substantive institutional revision. This dimension draws on deliberative democratic theory and the literature on algorithmic accountability, both of which treat legitimacy as inseparable from revisability. A system scores higher on Q when disagreement can be institutionally expressed and translated into change rather than ignored or averaged away.

The aggregate epistemic sovereignty index is then:

$$E = \eta_T T + \eta_A A + \eta_Q Q, \quad \sum_i \eta_i = 1 \quad (4)$$

This formulation clarifies why many contemporary sovereign AI projects remain epistemically thin despite substantial infrastructural investment. Systems may achieve local deployment, domestic branding, or improved performance in local languages while still lacking thick evaluation, local authorship over standards, or meaningful avenues of contestation. In such cases, operational sovereignty may increase while epistemic sovereignty remains limited.

A natural question is why our formalization of E does not directly incorporate model-output evaluation, given that output-based studies supply much of the evidence for the epistemic gap in Section 4. The asymmetry is deliberate. Output evaluations provide *ex post* evidence that epistemic sovereignty is absent, but they are poor constitutive measures of its presence: benchmark-style output testing is itself the kind of “thin” evaluation whose limits Section 4.2 documents, and a model could pass any fixed battery of cultural benchmarks while the authority to define, interpret, and revise those benchmarks remained external. We therefore define E in terms of the institutional conditions—thick evaluation, local authorship, contestability—under which output evaluation becomes epistemically meaningful. Output evaluation remains an essential diagnostic signal, particularly at the deployment stage of Table 1, but it enters the framework as evidence to be weighed within locally governed evaluative institutions, not as the definition of sovereignty itself.

Participatory legitimacy. Participatory legitimacy is not an external desideratum appended to sovereignty but a constitutive condition of it. Sovereignty, in the sense invoked by the political rhetoric of sovereign AI, is a claim to *collective self-determination*—authority exercised on behalf of and authorized by a political community. Stripped of the participatory dimension, what remains is state autonomy or state capacity: the ability of a government to act without

external interference, which is conceptually distinct from a community’s ability to shape the systems that govern its public life. As argued in Section 3.2, sovereign AI claims routinely invoke collective self-determination while being operationalized through narrow state-vendor pipelines that exclude many of the publics in whose name sovereignty is asserted. Control over infrastructure or legal jurisdiction does not, by itself, establish that the resulting authority is the community’s rather than merely the state’s. When AI systems mediate language access, education, welfare delivery, healthcare, or public administration, decisions about procurement, training data, alignment criteria, evaluation metrics, and acceptable model behavior become exercises in public value ordering. Participatory legitimacy therefore concerns whether those subject to these systems have meaningful standing in shaping them.

To formalize this dimension parsimoniously, we define participatory legitimacy over three core components: inclusiveness (Π), representational fairness (Λ), and deliberative power (Ω). Together, these capture whether participation is broad, equitable, and consequential rather than merely symbolic.

Inclusiveness (Π) refers to the extent to which materially affected groups are present within governance, consultation, and oversight processes. This includes end users in the aggregate, as well as linguistic minorities, Indigenous communities, civil society organizations, domain workers, and other populations differentially affected by deployment. The concept draws on democratic theories of inclusion, which hold that those subject to collective decisions should have opportunities to participate in shaping them. A system scores higher on Π when participation extends beyond state agencies, vendors, and technical elites to encompass the wider range of affected publics.

Representational fairness (Λ) concerns whether participation is distributed in ways that reflect social diversity rather than concentrated among already powerful actors. Formal inclusion alone may still reproduce elite capture if consultation bodies are dominated by politically connected groups, majority populations, or actors with disproportionate technical and institutional resources. This dimension draws on theories of political representation and critiques of technocracy, emphasizing that presence must be accompanied by equitable voice (Malgieri 2025). A system scores higher on Λ when marginalized or vulnerable communities are not merely invited but meaningfully represented relative to their stake in the system’s consequences.

Deliberative power (Ω) measures whether participation can actually shape outcomes. Consultation processes often collect input without altering procurement choices, benchmark design, deployment conditions, or redress mechanisms. By contrast, deliberative democratic theory treats legitimacy as dependent not only on speaking opportunities but on whether public reasoning can influence authoritative decisions. A system scores higher on Ω when participatory forums possess agenda-setting capacity, when feedback can trigger revision, and when disagreement is institutionally translated into change rather than absorbed as performative consultation.

Therefore, the aggregate participatory legitimacy index can be written as:

$$P = \rho_{\Pi}\Pi + \rho_{\Lambda}\Lambda + \rho_{\Omega}\Omega, \quad \sum_i \rho_i = 1 \quad (5)$$

This expression explains why many sovereign AI initiatives remain democratically thin despite strong claims of national control. Systems may be domestically procured, locally hosted, or publicly branded while affected communities remain excluded from consequential decisions. In such cases, operational sovereignty may increase while participatory legitimacy remains limited. The result is a model of sovereignty exercised for a population without being substantively authorized by it.

A note on measurement. The weighted-sum expressions above are ordinal, heuristic aggregators rather than cardinal measurement instruments; our case assessments accordingly assign qualitative levels (low, moderate, high) rather than point values. We deliberately do not prescribe weights. How data autonomy should trade off against model autonomy, for instance, depends on a country’s threat model, fiscal capacity, and strategic priorities—and, more fundamentally, who sets the weights is itself a question of participatory legitimacy: a weighting fixed by vendors or ministries alone would replicate at the meta-level the exclusion the framework is designed to diagnose. What the formalization fixes is simply a general structure: sovereignty has three non-substitutable dimensions, each with identifiable components, and the bottleneck condition holds under any weighting. Developing validated indicators and weight-elicitation procedures is a task for future empirical work.

Applying the framework to ALLaM and SEA-LION. Re-read through this framework, both initiatives score relatively high on *I*—domestic or regional jurisdictional control with substantial local compute and data—though reliance on foreign base architectures and toolchains constrains *M*. Both fall short on *E*: ALLaM across all three components (thin evaluation, centralized authorship, no contestation mechanisms), SEA-LION marginally better on *T* and *A* through SEA-HELM’s regional contributors, though involvement remains consultative and contestability is similarly absent. *P* is the weakest dimension in both cases: inclusiveness extends mainly to state actors and technical experts, and deliberative power is effectively absent. Under the bottleneck condition $S^{AI} = \min\{I, E, P\}$, each system’s overall sovereignty is capped by its lowest-scoring pillar despite substantial infrastructural achievement—precisely the configuration we identify as sovereignty theater: high *I* masking low *E* and *P*.

5.2 Sovereignty Theater and Its Consequences

The preceding analysis reveals a consistent pattern across the commercial ecosystem and the case studies: contemporary sovereign AI initiatives reliably achieve infrastructural sovereignty while systematically falling short on epistemic sovereignty—satisfying the visible, politically legible markers of sovereignty (local hosting, national branding, linguistic localization) while leaving the deeper struc-

tures of knowledge production, representation, and evaluation largely unchanged.

Two structural drivers underpin this phenomenon. First, the commodity form of sovereign AI requires standardization. Commercial vendors must produce scalable, modular solutions that can be deployed across jurisdictions. This standardization favors shared architectures, common training pipelines, and uniform evaluation metrics. Cultural specificity, particularly in its “thick,” context-dependent sense, resists such standardization. As a result, epistemic diversity is systematically compressed into forms that can be operationalized within generalized technical frameworks.

Second, the procurement model through which sovereign AI is acquired channels decision-making through state-vendor relationships. As discussed in Section 3.2, this institutional structure excludes the populations whose values are ostensibly being represented. Decisions about model selection, training data, alignment criteria, and evaluation benchmarks are made within a narrow pipeline of government actors and private firms. This reinforces both the epistemic gap (by limiting sources of knowledge) and the democratic deficit (by limiting participation).

The consequences of sovereignty theater are not merely conceptual. They manifest in downstream harms, particularly when sovereign AI systems are deployed in public-facing domains such as healthcare, welfare provision, education, and legal decision-making. These are not hypothetical domains. ALLaM already powers a national consumer chatbot and is offered to Saudi government entities through IBM’s watsonx platform on the DEEM sovereign cloud (IBM 2024), while SEA-LION serves as the base for Sahabat-AI, an Indonesian-language model family developed by GoTo and Indosat Ooredoo Hutchison and being integrated into consumer-facing services (AI Singapore 2025). In these contexts, models that appear culturally localized, because they operate in the local language or are hosted domestically, may nonetheless embed foreign normative assumptions in subtle ways. Policy recommendations, risk assessments, or classifications may reflect implicit priorities derived from Western training distributions, reshaping local decision-making processes without overt visibility.

This is precisely what makes the problem difficult to detect: the system passes a superficial sovereignty test—it looks and feels local—while failing a deeper epistemic one. The result is a form of normative drift, in which external value systems are gradually internalized within domestic governance infrastructures under the guise of sovereign control. Within our framework, this is a systematic divergence between *I* and *E*, compounded by low *P*: because sovereignty is bottlenecked by its weakest component, such systems remain operationally sovereign but epistemically dependent, legitimized through the language of national autonomy.

5.3 Toward Genuine Epistemic Sovereignty

If current sovereign AI initiatives fall short because they treat sovereignty as primarily an infrastructural problem, then achieving genuine epistemic sovereignty requires a shift in both evaluation and design. Rather than asking

whether systems are locally hosted or legally controlled, the relevant question becomes whether they meaningfully reflect the knowledge systems, values, and interpretive frameworks of the populations they serve. A sovereign AI initiative demonstrates epistemic sovereignty to the extent that it satisfies several core criteria.

First, multistakeholder participation is essential. Epistemic sovereignty requires that affected communities play a role in defining what counts as appropriate, accurate, or legitimate model behavior. This goes beyond consultation or data contribution; it entails participation in the design of alignment processes, evaluation criteria, and governance structures. Without such participation, the *P* dimension remains underdeveloped, constraining overall sovereignty regardless of infrastructural capacity. A growing literature on participatory AI offers both concrete methods and cautions here: co-design and community-governed data practices provide templates for consequential participation, while critics warn that shallow consultation can function as a legitimating “design fix” (Birhane et al. 2022; Delgado et al. 2023; Sloane et al. 2022)—precisely the participatory analogue of sovereignty theater.

Second, systems must rely on culturally grounded data generation, rather than merely adapting existing web-scale corpora. As the evidence in Section 4 demonstrates, reliance on scraped or translated data reproduces existing epistemic hierarchies. Achieving epistemic sovereignty requires active investment in building datasets that reflect local knowledge systems, linguistic practices, and cultural contexts, particularly for underrepresented or low-resource communities.

Third, alignment processes must be demographically and culturally representative. Current alignment practices, often based on geographically concentrated annotator pools, introduce systematic biases into model behavior. Expanding the diversity of annotators and incorporating plural perspectives into reinforcement learning and evaluation processes is necessary to avoid encoding narrow interpretations of appropriateness or correctness.

Fourth, sovereign AI systems should incorporate ongoing cultural auditing mechanisms post-deployment. Epistemic alignment is not a one-time achievement but a continuous process, especially in dynamic social contexts. Regular auditing, potentially involving independent institutions or community oversight, can help detect and correct forms of misalignment that emerge over time.

Fifth, transparency regarding model provenance and limitations is critical. Governments and vendors should clearly disclose the origins of base models, the extent of fine-tuning, and the limits of epistemic adaptation. Equally important is transparency about evaluation benchmarks: what standards are used to assess cultural alignment, and whose definitions of success those standards reflect. Without such transparency, claims of sovereignty remain difficult to substantiate or contest.

Table 1 operationalizes our framework as a diagnostic tool for policymakers and procurement teams. The matrix specifies decision points across four lifecycle stages—procurement, pre-development, development, and deployment—and asks, at each stage, what would have to be

true for the *I*, *E*, and *P* dimensions to be substantively secured rather than nominally claimed. The questions reiterate that sovereignty operates as a bottleneck condition: deficits in epistemic or participatory legitimacy cannot be offset by infrastructural control alone. This tool offers a diagnostic for distinguishing genuine sovereignty from sovereignty theater in practice.

A qualification concerns the status of these criteria. The components of *E* and *P* should be understood as ideal-typical reference points rather than thresholds that any single initiative is likely to fully achieve. In the current political economy of AI—characterized by the concentration of frontier capabilities in a small number of U.S. and Chinese firms, the structural dependence of even well-resourced sovereign initiatives on foreign base models and hardware, and procurement cycles that favor turnkey commercial packages—thick evaluation, locally grounded epistemic authorship, and meaningful deliberative power face substantial structural headwinds. Our claim is not that any sovereign AI program can be expected to score uniformly high across all dimensions, but that the framework makes visible where particular initiatives fall short and why infrastructural achievement alone should not be read as sovereignty in the fuller sense the rhetoric invokes. Treating the criteria as ideal types also clarifies the analytic work of the bottleneck condition: it identifies the binding constraint on overall sovereignty, which in the current landscape is almost always *E* or *P* rather than *I*.

Even with such a diagnostic in hand, however, pursuing epistemic sovereignty raises its own normative tensions. In particular, aligning models with “local values” may conflict with commitments to minority rights, gender equality, or other universalist principles. A system that faithfully reflects dominant cultural norms within a society may reproduce or even amplify forms of exclusion or hierarchy. This presents a genuine dilemma: epistemic sovereignty may require deference to locally grounded value systems, yet those systems may themselves be contested or normatively problematic.

This tension highlights the need for further research at the intersection of AI governance, political theory, and ethics. The challenge is to determine how plural, contested, and evolving value systems can be represented in ways that are both legitimate and normatively defensible. A related dilemma, flagged earlier in Section 3.2, concerns inclusion itself: representing marginalized communities in sovereign AI systems may require forms of data collection that expose those same communities to surveillance or reidentification risks (Nissenbaum 2004). Both tensions—between local values and universalist commitments, and between representation and protection—warrant sustained attention in future work on participatory AI governance.

What the present analysis establishes, however, is a more immediate point. Sovereign AI, as currently practiced, does not resolve the problem it is meant to address. Without confronting the epistemic and democratic dimensions of sovereignty, efforts to localize AI may simply end up reproducing global asymmetries.

Dimension	Procurement	Pre-Development (Planning & Design / Data Collection)	Development (Training / Evaluation)	Deployment (Application / Oversight)
Infrastructural Sovereignty $I = I(D, C, M)$ <i>Data, compute, model autonomy</i>	Do we have transparent visibility into the model’s provenance and modification limits?	Does the system design ensure long-term control over infrastructure, compute, and data storage?	Who controls the evaluation and benchmarking infrastructure, and are we dependent on externally defined performance standards?	Can we audit, modify, and/or replace the model, or are we constrained by external infrastructure dependencies?
Epistemic Sovereignty $E = E(T, A, Q)$ <i>Thick evaluation, authorship, contestability</i>	What cultural assumptions are embedded in the system’s pretraining, and can they be changed downstream?	Do training data and model behavior reflect locally defined cultural knowledge and values?	Are we using thick evaluation approaches that capture cultural nuance?	Do system outputs accurately reflect local knowledge and values, and can epistemic failures be identified, recorded, and contested over time?
Participatory Legitimacy $P = P(\Pi, A, \Omega)$ <i>Inclusiveness, representational fairness, deliberative power</i>	Were affected communities included in vendor selection, and did they influence the decision?	Do affected groups have power over how their data is collected, used, and interpreted?	Who decides alignment policy and the criteria for evaluating cultural fit?	Do affected groups have deliberative power to influence system changes, or is participation symbolic?

Table 1: Sovereign AI Lifecycle Decision Matrix — Mapping the three sovereignty dimensions to four lifecycle stages.

6 Conclusion

Sovereign AI responds to a genuine problem. As the evidence reviewed in this paper confirms, mainstream LLMs carry systematic cultural biases rooted in their predominantly Western training data, architectures, and alignment processes. The aspiration to build AI systems that reflect local languages, values, and knowledge traditions is both reasonable and important. But the paper has shown that sovereign AI, as currently commercialized and procured, does not deliver on this aspiration. The epistemic gap persists because of how the technology is built (upstream assumptions survive downstream fine-tuning) and how it is acquired—procurement pipelines channel decisions through state-vendor relationships that exclude the populations whose values are ostensibly at stake.

This paper makes three contributions. First, it bridges two literatures that have developed largely in parallel: scholarship on sovereign AI and digital sovereignty as matters of geopolitics and policy, and empirical research on cultural misalignment in LLMs. Connecting these bodies of work reveals that the technical limitations documented by the latter directly undermine the political promises made by the former. Second, the paper introduces the distinction between operational and epistemic sovereignty and develops the concept of sovereignty theater as an analytical lens—applied here to the commercial vendor ecosystem and to the cases of ALLaM and SEA-LION, but with reach beyond sovereign AI: analogous gaps between performed and substantive autonomy arise wherever states assert control over technologies, from sovereign clouds to semiconductor supply chains to digital public infrastructure, whose epistemic and material foundations they do not govern. Third, our formal model provides a framework for understanding why infrastructural investment alone cannot achieve genuine sovereignty. We operationalize this framework as a lifecycle decision matrix

(Table 1) that future work can apply to assess specific national initiatives.

Several limitations should be acknowledged. The case studies of ALLaM and SEA-LION rely on publicly available documentation and published technical papers; internal deliberation processes and procurement negotiations remain opaque. We also do not directly evaluate the outputs of these models: a more comprehensive and systematic evaluation of sovereign models is a natural next step. Future research should also pursue interview-based work with sovereign AI developers, procurement officials, and affected communities; comparative analysis of procurement governance across a broader range of countries and vendor configurations; deeper engagement with Indigenous data governance frameworks such as the CARE principles (Carroll et al. 2020) as models of community epistemic authority; and the question of what follows if epistemic sovereignty conflicts with commitments to minority rights, gender equality, or democratic pluralism.

The implication of this analysis is not that sovereign AI is never worth pursuing. The underlying concerns about cultural and epistemic dependency on foreign AI systems are legitimate, and the alternative—uncritical reliance on a small number of globally dominant models—carries its own risks. But sovereign AI must be reoriented from a product that can be purchased to a political project that must be constructed. The question policymakers should be asking is not just “where is our AI located?” but “whose values does it reflect, and who decided?” Ultimately, the AI-ethics community is well-positioned to contribute to this reorientation by developing culturally grounded evaluation frameworks, designing participatory governance models for AI procurement, and producing the empirical evidence needed to hold sovereign AI initiatives accountable to the populations they claim to serve.

Acknowledgments

We are grateful for helpful feedback from Tendayi Achiume, Andrew Grotto, Florence G'sell, Caroline Meinhardt, Peter Norvig, Julian Nyarko, Juan Pava, Drew Spence, Quinn Waless, others at the Stanford Institute for Human-Centered AI (HAI), and participants of Stanford's Annual Tech and Ethics Research Conference.

References

- Agarwal, D.; Shukla, A.; Sitaram, S.; and Vashistha, A. 2025. Fluent but Foreign: Even Regional LLMs Lack Cultural Alignment. arXiv:2505.21548.
- AI Singapore. 2025. Sahabat AI: Case Study. <https://sealion.ai/case-study/sahabat-ai/>.
- Alrajjal, K. 2025. Saudi Arabia is Building an AI Powerhouse. CNN, November 2, 2025. <https://www.cnn.com/2025/11/02/tech/saudi-arabia-ai-powerhouse>. Accessed: 2026-04-12.
- Arora, A.; Kaffee, L.-A.; and Augenstein, I. 2022. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. arXiv:2203.13722.
- Bellanova, R.; Carrapico, H.; and Duez, D. 2022. Digital/Sovereignty and European Security Integration: An Introduction. *European Security*, 31(3): 337–355.
- Benhabib, S. 1994. Deliberative Rationality and Models of Democratic Legitimacy. *Constellations*, 1(1): 26–52.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.
- Buyl, M.; Rogiers, A.; Noels, S.; Bied, G.; Dominguez-Catena, I.; Heiter, E.; Johary, I.; Mara, A.-C.; Romero, R.; Lijffijt, J.; and De Bie, T. 2024. Large Language Models Reflect the Ideology of Their Creators. arXiv:2410.18417.
- Carroll, S. R.; Garba, I.; Figueroa-Rodríguez, O. L.; Holbrook, J.; Lovett, R.; Materechera, S.; Parsons, M.; Raseroka, K.; Rodriguez-Lonebear, D.; Rowe, R.; et al. 2020. The CARE principles for indigenous data governance. *Data science journal*, 19(1): 43–43.
- Chavez, P. 2026. The Sovereignty Gap in U.S. AI Statecraft. <https://www.lawfaremedia.org/article/the-sovereignty-gap-in-u.s.-ai-statecraft>. Lawfare, February 16, 2026. Accessed: 2026-04-12.
- Chavez, P.; Chilukuri, V.; and Scanlon, R. 2026. Sovereign AI Index: Tracking the Global Push for AI Self-Reliance. Center for a New American Security, April 2026. <https://interactives.cnas.org/reports/sovereign-ai-index/>. Accessed: 2026-05-19.
- Couture, S.; and Toupin, S. 2019. What Does the Notion of “Sovereignty” Mean When Referring to the Digital? *New Media & Society*, 21(10): 2305–2322.
- Dale, R. 2025. Sovereign AI in 2025. *Natural Language Processing*, 31(5): 1312–1321.
- Dawson, F.; Mosunmola, Z.; Pocker, S.; Dandekar, R. A.; Dandekar, R.; and Panat, S. 2024. Evaluating Cultural Awareness of LLMs for Yoruba, Malayalam, and English. arXiv:2410.01811.
- de Ocariz Borde, H. S. 2025. Sovereign AI vs AI Polyglots. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5639470. SSRN Working Paper, 44 pages. Accessed 2026-05-19.
- Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in AI design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.
- Dell Technologies. 2025. Empowering Sovereign AI with Cultural Relevance: NAVER Cloud Case Study. <https://www.delltechnologies.com/asset/en-us/solutions/business-solutions/customer-stories-case-studies/naver-cloud-case-study.pdf>. Accessed: 2026-04-15.
- Dryzek, J. S. 2002. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford: Oxford University Press. ISBN 9780199250431.
- European Commission, Directorate-General for Translation. 2025. An EU Large Language Model to Promote Language Equality. https://translation.ec.europa.eu/language-data-and-ai-using-ai-break-down-language-barriers/eu-large-language-model-promote-language-equality_en. Accessed: 2026-04-15.
- G'sell, F. 2025. Digital Authoritarianism: From State Control to Algorithmic Despotism. In De Gregorio, G.; Pollicino, O.; and Valcke, P., eds., *The Oxford Handbook of Digital Constitutionalism*. Oxford University Press. Online edition. Accessed: 2026-05-21.
- Gutmann, A.; and Thompson, D. 2000. Why deliberative democracy is different. *Social Philosophy and Policy*, 17(1): 161–180.
- Hewlett Packard Enterprise. 2025. HPE Advances Government and Enterprise AI Adoption Through Secure AI Factory Innovations With NVIDIA. <https://www.hpe.com/us/en/newsroom/press-release/2025/10/hpe-advances-government-and-enterprise-ai-adoption-through-secure-ai-factory-innovations-with-nvidia.html>. HPE Press Release, October 2025. Accessed: 2026-04-15.
- Humain AI. 2024. ALLaM-7B-Instruct-preview. Hugging Face model repository. <https://huggingface.co/humain-ai/ALLaM-7B-Instruct-preview>. Accessed: 2026-04-12.
- IBM. 2024. IBM watsonx.ai and ALLaM Available to Saudi Government Entities with DEEM Cloud. <https://mea.newsroom.ibm.com/watsonx-and-ALLaM-on-DEEM-cloud>.
- Kaminski, M. E.; and Maltgieri, G. 2025. Impacted Stakeholder Participation in AI and Data Governance. *Yale Journal of Law & Technology*, 27: 247.
- Kukutai, T.; and Taylor, J., eds. 2016. *Indigenous Data Sovereignty: Toward an Agenda*. CAEPR Research Monograph No. 38. Canberra: ANU Press.

- Malgieri, G. 2025. Scalable Fairness. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25. Forthcoming.
- Masoud, R. I.; Liu, Z.; Ferianc, M.; Treleaven, P.; and Rodrigues, M. 2023. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. arXiv:2309.12342.
- Meinhardt, C.; Pava, J. N.; Yee, C.; and Landay, J. A. 2026. The Commercial Landscape of AI Sovereignty Offerings. Stanford Institute for Human-Centered Artificial Intelligence (HAI) Issue Brief, July 2026. <https://hai.stanford.edu/policy/the-commercial-landscape-of-ai-sovereignty-offerings>. Accessed: 2026-07-19.
- Metcalfe, T. 2026. AI Safety and Regulatory Capture. *AI & Society*, 41: 2451–2466.
- Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33: 659–684.
- Ng, R.; Nguyen, T. N.; Huang, Y.; Tai, N. C.; Leong, W. Y.; Leong, W. Q.; Yong, X.; Ngui, J. G.; Susanto, Y.; Cheng, N.; Rengarajan, H.; Limkonchotiwat, P.; Hulagadri, A. V.; Teng, K. W.; Tong, Y. Y.; Siow, B.; Teo, W. Y.; Lau, W.; Tan, C. M.; Ong, B.; Ong, Z. H.; Montalan, J. R.; Chan, A.; Antonyrex, S.; Lee, R.; Choa, E.; Tat-Wee, D. O.; Liu, B. J. D.; Tjhi, W. C.; Cambria, E.; and Teo, L. 2025. SEA-LION: Southeast Asian Languages in One Network. Accepted at IJCNLP-AAACL 2025 (Main Track), arXiv:2504.05747.
- Nissenbaum, H. 2004. Privacy as contextual integrity. *Wash. L. Rev.*, 79: 119.
- NVIDIA Corporation. 2025. Telcos Across Five Continents Are Building NVIDIA-Powered Sovereign AI Infrastructure. <https://developer.nvidia.com/blog/telcos-across-five-continents-are-building-nvidia-powered-sovereign-ai-infrastructure/>. NVIDIA Technical Blog. Accessed: 2026-04-15.
- NVIDIA Corporation. 2026a. AI Factories for Telecommunications. <https://www.nvidia.com/en-us/industries/telecommunications/ai-factories/>. Accessed: 2026-04-15.
- NVIDIA Corporation. 2026b. What Is an AI Factory? <https://www.nvidia.com/en-us/glossary/ai-factory/>. NVIDIA Glossary. Accessed: 2026-04-15.
- Oracle Corporation. 2026. Sovereign AI. <https://www.oracle.com/artificial-intelligence/sovereign-ai/>. Accessed: 2026-04-15.
- Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrana, F. G.; Song, I.; Oh, A.; and Augenstein, I. 2025. Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, 51(3): 907–1004.
- Phillips, D. 2026. Microsoft Sovereign Cloud Adds Governance, Productivity and Support for Large AI Models Securely Running Even When Completely Disconnected. <https://blogs.microsoft.com/blog/2026/02/24/microsoft-sovereign-cloud-adds-governance-productivity-and-support-for-large-ai-models-securely-running-even-when-completely-disconnected/>. The Official Microsoft Blog, February 24, 2026. Accessed: 2026-04-15.
- Png, M.-T. 2022. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1434–1445. New York, NY, USA: Association for Computing Machinery.
- Press Information Bureau. 2026. Sarvam AI Powering a Made-in-India AI Revolution: Advancing India's Sovereign, Multilingual AI Ecosystem for Public Good. <https://www.pib.gov.in/PressReleasePage.aspx?PRID=2231169®=3&lang=1>. Government of India, Ministry of Information and Broadcasting, Explainer ID 157508, February 21, 2026. Accessed: 2026-04-15.
- Qadri, R.; Diaz, M.; Wang, D.; and Madaio, M. 2025. The Case for “Thick Evaluations” of Cultural Representation in AI. arXiv:2503.19075.
- Roberts, H. 2024. Digital sovereignty and artificial intelligence: a normative approach. *Ethics and Information Technology*, 26(4): 70.
- Singh, S. K.; and Sengupta, S. 2025. Sovereign AI: Rethinking Autonomy in the Age of Global Interdependence. arXiv:2511.15734.
- Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–6.
- Sukumar, A. M. 2025. Tech's ‘Sovereignty Washing’ in Europe Will Ripple in the Global South. Lawfare, October 9, 2025. <https://www.lawfaremedia.org/article/tech-s-sovereignty-washing--in-europe-will-ripple-in-the-global-south>. Accessed: 2026-05-08.
- Susanto, Y.; Hulagadri, A. V.; Montalan, J. R.; Ngui, J. G.; Yong, X.; Leong, W. Q.; Rengarajan, H.; Limkonchotiwat, P.; Mai, Y.; and Tjhi, W. C. 2025. SEA-HELM: Southeast Asian Holistic Evaluation of Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.
- Taeihagh, A. 2021. Governance of artificial intelligence. *Policy and Society*, 40(2): 137–157.
- Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus*, 3(9): 346.
- Ustun, A.; Tournesac, A.; Bennici, L.; Takkar, K.; and Drahmoune, N. 2025. The Sovereign AI Agenda: Moving from Ambition to Reality. <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/the-sovereign-ai-agenda-moving-from-ambition-to-reality>. McKinsey & Company, Tech Forward, December 18, 2025. Accessed: 2026-04-15.
- Varela Sandoval, F. J.; Wilkinson, I.; Krasodomski-Jones, A.; and Wilkinson, R. 2026. How Middle Powers Can Weather US and Chinese AI Dominance. Chatham House Research Paper, February 2026. <https://www.chathamhouse.org/2026/02/how-middle-powers-can-weather-us-and-chinese-ai-dominance>. Accessed: 2026-05-19.