Linguistic Diversity, Artificial Intelligence, and the Governance of the Infosphere:
Risk Narratives, Interpretive Frameworks, and the Limits of the 'Technical Fix'

Paola Cantarini[1]

Abstract

Artificial intelligence systems — especially large language models — have quietly transformed language into a form of infrastructure. Yet the dominant public and policy discourse on linguistic diversity in AI remains captured by a persistent narrative: that the problem is one of scale, and that the solution is technical. This article analyses that narrative as a socially stabilized pattern of meaning — a risk narrative — that actively shapes how the problem is defined, which risks are rendered visible, and which governance responses are foreclosed. Drawing on the concepts of interpretive frameworks (Deutungsmuster) and narratives as collective sense-making structures, the article shows that the 'technical fix' narrative does not merely misdiagnose the problem; instead, it reproduces the very power asymmetries it claims to address. An alternative framing is developed, grounded in information ethics, commons theory, and Indigenous data sovereignty, which reconceptualizes linguistic data portals as data trusts and data commons. This reframing is not merely a governance proposal. It is a counter-narrative: one that renders visible the structural and political dimensions of linguistic inequality in AI, and opens space for accountability, collective rights, and democratic agency — including within the European fundamental rights framework that provides the normative backbone of data protection law.

1 The analytical framework and policy proposals advanced in this article — including the Linguistic Diversity Impact Index (LDII) and data trust models for linguistic sovereignty — were developed by the author in the context of dialogues with Brazil's Ministry of Foreign Affairs (Itamaraty) and other institutional stakeholders. This article constitutes an independent scholarly work. The views, arguments, and conclusions expressed herein are solely those of the author and do not necessarily reflect the official positions of the Government of Brazil or Itamaraty.

1. Introduction: A Problem and Its Dominant Story

Over the past decade, language has ceased to be merely an object processed by artificial intelligence. It has become one of its core substrates. In contemporary AI systems — particularly large language models and generative architectures — linguistic data functions as infrastructure in the strict sense: it enables, constrains, and structures what systems can do, what they can say, and, indirectly, what people can access. This shift has unfolded gradually. Translation systems, conversational agents, automated decision-support tools, and public-sector digital services accumulated quietly. Taken together, they reveal a profound transformation: linguistic data now underwrites administrative capacity, mediates access to rights and services, and increasingly shapes the conditions under which individuals and communities participate in digital societies.

What is at stake here is not simply representation. It is power.

Yet the dominant way of framing this problem — in policy documents, technical roadmaps, and international governance fora — tells a different story. In this story, linguistic inequality in AI is a problem of underrepresentation: too few datasets, too little training data, too many 'low-resource languages.' The proposed solutions follow a predictable pattern: expand corpora, improve benchmarks, scale multilingual models. Progress is measured in coverage statistics. Risk is understood as technical inadequacy.

The solution is more data, better engineering.

This article takes that story seriously — as an object of analysis. The 'scale-and-fix' account is not merely a policy preference. It is a socially stabilized narrative: a pattern of meaning that links actors, events, and causalities into a coherent sense-making structure, shapes how risk is perceived, and determines which interventions appear legitimate (Müller-Funk 2002; Amlinger 2023). As such, it does not merely describe the problem of linguistic diversity in AI. It actively constitutes it — in ways that systematically foreclose the governance questions that matter most. The stakes of this foreclosure extend directly to the normative foundations of data protection law. In European constitutional and supranational law, the protection of personal data is enshrined as a fundamental right under Article 8 of the Charter of Fundamental Rights of the European Union (CFR). The Court of Justice of the European Union (CJEU) has consistently interpreted this right not as a merely procedural guarantee, but as a substantive condition for the exercise of other fundamental rights — dignity, expression, non-discrimination. Linguistic data, as this article argues, sits squarely within

this framework: it is personal data, it carries identity, and its extractive governance threatens the very rights the GDPR and the CFR are designed to protect. A risk narrative that frames linguistic inequality as a coverage problem renders this fundamental rights dimension invisible, depoliticizing what is in fact a constitutional question. This article argues for a narrative reorientation that makes that question visible — and actionable.

The central argument is therefore twofold. First, the dominant narrative of linguistic diversity in AI misrepresents the nature of the problem, rendering structural and political risks invisible while presenting extractive data practices as solutions. Second, an alternative narrative is both possible and necessary — one grounded in governance, collective rights, and institutional accountability. Developing that alternative requires not only different policy proposals but a different interpretive framework: a different way of perceiving what is at risk, who bears responsibility, and what counts as a solution.

The article proceeds as follows. Section 2 analyses the dominant narrative and the interpretive framework that sustains it, applying Bacchi's (2009) 'What's the problem represented to be?' method to expose what the 'technical fix' approach misdiagnoses and conceals. Section 3 demonstrates where this narrative fails — the risks it obscures, the harms it enables. Section 4 develops an alternative framing grounded in commons theory, fiduciary stewardship, and Indigenous data sovereignty, and analyses the obstacles to its circulation. Section 5 draws conclusions about what it would mean to govern linguistic data as a condition of informational sovereignty and fundamental rights protection.

Methodological note. This article combines a normative institutional analysis of the AI governance field with a narrative-analytical lens. It draws on publicly available policy documents, governance debates, and institutional experiences to reconstruct the dominant interpretive frameworks and narratives operative in the field. The aim is conceptual clarification of how problems are framed and what those framings do. This article does not employ narrative accounts in the methodological sense established by Labov and Waletzky (1973) and developed in qualitative social research by Küsters (2022), where narration refers to the elicitation and analysis of temporally ordered first-person accounts of lived experience. The use of 'narrative' here follows instead the sociological and discourse-analytical tradition of Müller-Funk (2002) and Amlinger (2023), which treats narratives as socially stabilized patterns of collective meaning — structures that circulate beyond any individual account and shape public problem representations.

2. The Dominant Narrative: 'Scale and Fix'

2.1 The Structure of the Narrative

In the sociology of knowledge, narratives are distinguished from individual accounts of experience. They are socially and culturally stabilized patterns of meaning that condense underlying interpretive frameworks into communicable, often highly compressed forms — metaphors, storylines, problem definitions (Müller-Funk 2002; Amlinger 2023). They do not merely represent realities; they actively shape what becomes visible as a problem, who is recognized as a legitimate actor, and which responses appear rational.

The dominant narrative of linguistic diversity in AI can be summarized in a single storyline: language diversity is a coverage problem, and coverage problems have technical solutions. In this storyline, the actors are engineers and data scientists; the problem is insufficient training data; the causal mechanism is scale; and the solution is more data, better models. Risk is located in technical inadequacy. Progress is measured in benchmark scores and language coverage statistics.

This narrative draws its plausibility from an underlying interpretive framework — a Deutungsmuster in the sense of Siebert (1996) and Arnold (1985, 1991). Interpretive frameworks are relatively stable, often implicit cognitive schemas through which actors perceive their environment, reduce complexity, and orient action. They are shaped biographically and institutionally. The interpretive framework operative here might be described as 'technological solutionism' (Morozov/Bria 2018): the assumption that complex social problems are, at their core, engineering problems, and that sufficient technical ingenuity will resolve them. Within this framework, governance appears at best as a secondary concern — important for implementation, but not constitutive of the problem itself.

This framework does not emerge from individual misunderstanding. It is institutionally embedded. Funding streams, publication incentives, procurement frameworks, and international standard-setting processes all reward technical outputs — datasets, benchmarks, model architectures — over governance innovations. The actors who dominate AI governance debates — technology companies, computer science departments, large multilateral organizations — share this interpretive orientation, and reproduce it through the narratives they disseminate.

2.2 How the Narrative Travels

Risk narratives, as Amlinger (2023) and Müller-Funk (2002) emphasize, do not merely describe risks — they produce risk perceptions. The 'scale-and-fix' narrative is particularly effective because it resonates with several widely shared interpretive frameworks simultaneously: the narrative of scientific progress through accumulation of data; the narrative of market efficiency as the appropriate mechanism for resource allocation; and the narrative of technology as neutral infrastructure, separable from the social relations in which it is embedded.

In international AI governance fora, this narrative travels efficiently. Delegations emphasize inclusivity and diversity as values, then default to scaling proposals as solutions. Governance is mentioned but rarely specified. The gap between normative commitments and institutional design is striking — and predictable, given the interpretive framework that structures the debate. 'Language diversity' becomes a symbolic resource, invoked to signal commitment, then quietly set aside when policy specifics are negotiated. The narrative also travels through technical documentation. Audit reports of training corpora consistently show that contemporary language models rely overwhelmingly on data from a very small number of high-resource languages, with English occupying a dominant position (Joshi et al. 2020; Kreutzer et al. 2022). These audits document the problem accurately. But they typically embed their findings within a framework that treats the imbalance as a gap to be filled rather than a structure to be challenged. Coverage becomes the metric of progress. Governance remains peripheral.

2.3 What the Narrative Renders Invisible: A Problem-Framing Analysis

Carol Bacchi's (2009) 'What's the Problem Represented to Be?' (WPR) method provides an analytical framework for examining precisely this kind of narrative operation. Rather than taking problem representations at face value, WPR asks: what does a given framing presuppose about the nature of the problem? What does it render invisible? And what subject positions and governance responses does it produce as natural or inevitable? Applying two of Bacchi's core questions to the 'scale-and-fix' narrative reveals its structural effects with precision.

First question: What is the problem represented to be? In the dominant narrative, the problem of linguistic diversity in AI is represented as a deficit

problem — a shortage of data, of coverage, of engineering attention. This representation presupposes that the informational environment for AI is in principle neutral and correctable through addition. It treats linguistic communities as passive sources of data rather than as rights-bearing actors with governance interests. It constructs the solution as technical before any question of institutional design has been asked. The subject produced by this framing is the data scientist who fills gaps, not the rights-holder who contests extraction.

Second question: What is left unproblematic? The 'scale-and-fix' framing leaves unproblematic the institutional arrangements through which linguistic data is collected, licensed, and controlled. It treats data availability as a neutral technical fact rather than as an artifact of licensing regimes, economic incentives, and historical patterns of digitization (Longpre et al. 2024). It does not ask who benefits from expanded corpora, or whether the communities whose languages are included retain any meaningful say over how their linguistic heritage is used. The question of control — the central governance question — is simply not represented as a problem. And what is not represented as a problem cannot generate governance responses.

Two further invisibilities follow from this analysis. The narrative obscures the epistemic consequences of linguistic homogenization: when automated systems are trained primarily on hegemonic linguistic categories, they encode particular worldviews as defaults that travel into downstream applications — public services, educational tools, content moderation systems. Speakers of underrepresented languages are not merely underserved; they are rendered partially invisible within computational systems that increasingly mediate access to rights and services. This is not a performance deficiency. It is an epistemic asymmetry with direct consequences for democratic participation and fundamental rights (Fricker 2007). The narrative also obscures surveillance and extraction risks specific to linguistic data: vocabulary choices, syntactic habits, and discourse structures can function as identifiers sufficient to distinguish individuals or groups, even when explicit identifiers are removed (Brennan et al. 2012; Zuboff 2019). For marginalized communities, linguistic visibility can translate directly into exposure — a risk that the 'scale-and-fix' frame cannot perceive because it has already constituted the data as a resource to be expanded, not a vulnerability to be governed.

3. Where the Narrative Fails: Structural Risks and Misspecified Solutions

3.1 Algorithmic Linguistic Racism as Structural Outcome

Research in raciolinguistics and critical algorithm studies has shown that AI systems often reproduce linguistic hierarchies aligned with race, class, and colonial histories (Rosa/Flores 2017; Noble 2018; Benjamin 2019). This phenomenon has been described as algorithmic linguistic racism. Dominant linguistic varieties are treated as normative. Vernaculars, Indigenous languages, and non-standard forms are misclassified, flagged as noise, or excluded altogether. The system does not announce this hierarchy. It enacts it. Crucially, these outcomes are not the result of isolated design errors or insufficient data. They emerge from the interaction between historical inequalities and contemporary data practices. Training data reflects existing power relations; models trained on such data inevitably inherit them. Linguistic AI systems function as infrastructures of power rather than neutral tools (Couldry/Mejias 2019; Kwet 2019).

The 'scale-and-fix' narrative misrepresents this dynamic as a gap-filling problem. It treats algorithmic linguistic racism as a residual technical imperfection, correctable through more data and better architecture. But what is being corrected — and what is being reinforced — depends entirely on whose linguistic categories structure the training process, whose standards define 'quality,' and whose institutions govern data curation. These are governance questions that the dominant narrative has already defined away. Scaling efforts that leave governance structures unchanged may actually intensify extractive dynamics: more data collected under the same institutional arrangements means more appropriation of linguistic resources from communities that retain no meaningful control over their use.

3.2 The 'Technical Fix' as a Risk Narrative in Itself

The section of the dominant discourse most directly relevant to this volume's focus on data protection, privacy, and surveillance concerns what might be called 'technical fix narratives.' In policy consultations, two technical measures are routinely proposed as solutions to the privacy risks of linguistic data: differential privacy and federated learning. These tools are presented not merely as technical options but as reassurances — as evi-

dence that the problem has been addressed and that affected communities are protected.

This framing deserves scrutiny as a risk narrative in its own right. Differential privacy introduces noise to protect individual records. In low-resource or endangered language contexts, this noise can erase precisely the features that make the data valuable for revitalization and cultural transmission. A technique that provides meaningful privacy protection for majority language data may render minority language data useless. The 'technical fix' narrative applies a single solution frame to structurally different contexts — and in doing so, hides context-specific harms behind generic reassurances. This is a narrative operation: it produces the perception of protection while foreclosing inquiry into whether that protection is real for the communities most at risk.

Federated learning reduces centralization but does not eliminate power asymmetries or opaque downstream uses. Distributing data processing across nodes does not determine who controls the model outputs, who benefits from their application, or whose values guide system design. It relocates the locus of data processing without restructuring the governance relationships that determine how value is distributed and risk is borne. From the perspective of data protection law, this matters: the GDPR's accountability principle (Article 5(2)) requires that data controllers be able to demonstrate compliance — a requirement that federated architectures can formally satisfy while leaving substantive accountability for community-level harms entirely unaddressed.

The deeper problem with technical fix narratives is that they perform a substitution: they replace governance decisions with technical choices, and in doing so, actively depoliticize what are irreducibly political questions. Technical safeguards cannot decide which uses of linguistic data are legitimate, how benefits should be shared, or whose communities count as beneficiaries. When technical narratives present themselves as substitutes for governance, they make those questions harder to raise and harder to contest. Communities that rely on differential privacy or federated learning as their primary protection, without institutional recourse, are exposed to harms that these tools structurally cannot address.

3.3 Floridi's Infosphere and the Temporal Dimension of Risk

Luciano Floridi's concept of the infosphere offers a framework for understanding a category of risk that the dominant narrative systematically neglects: the cumulative, slow-moving degradation of the informational environment (Floridi 2013, 2014). It is worth noting, before deploying this framework, that it has attracted criticism: the ecological analogy can naturalize the informational environment, obscuring the social relations and power asymmetries that produce it; and the infosphere concept risks presenting as a neutral environment what is in fact a contested political terrain shaped by specific actors with specific interests. These are legitimate objections, and they counsel against treating Floridi's framework as a complete account. Used carefully and in conjunction with the power-analytical perspectives of Couldry, Kwet, and Benjamin, however, the infosphere concept contributes a temporal dimension that the dominant narrative cannot supply.

In Floridi's account, contemporary societies are increasingly embedded in an infosphere composed of interactions between human and artificial agents. Language, within this environment, is the medium through which meaning is stabilized and contested. If the infosphere is understood as an environment, then linguistic diversity functions as a condition of its health — and linguistic homogenization in AI is not merely a cultural loss, but a form of informational degradation. Accuracy scores and benchmark performance do not capture what disappears when linguistic variety is compressed. They do not measure which concepts become harder to articulate, which epistemic traditions fade, or which forms of expression are quietly sidelined. This temporal dimension is precisely what the 'scale-and-fix' narrative cannot perceive. The narrative is oriented toward measurable, near-term improvements. Infospheric degradation does not announce itself as a crisis. It unfolds slowly, through defaults and design choices that appear reasonable in isolation. Linguistic diversity erodes not because anyone intends it, but because governance fails to keep pace with technological scaling. By the time the loss becomes visible, it may be difficult to reverse.

4. An Alternative Framing: Linguistic Data as Collectively Governed Commons

4.1 The Interpretive Shift Required

Developing an alternative to the dominant narrative requires more than different policy proposals. It requires a different interpretive framework: a different way of perceiving what is at risk, who bears responsibility, and what counts as a solution. The interpretive shift proposed here can be stated plainly: linguistic data should be understood not as a technical resource to be scaled, but as a collectively governed commons, subject to the same governance obligations that apply to other shared resources on which communities depend. This shift does not require abandoning the technical concerns that motivate the dominant narrative. Expanding linguistic coverage in AI systems matters. Improving model performance across languages matters. But these objectives become meaningful only within a governance framework capable of addressing the structural asymmetries, power relations, and collective rights that the 'scale-and-fix' narrative systematically obscures. Governance is not an add-on to the technical enterprise. It is its condition of legitimacy.

4.2 Commons and Trusts as Governance Frameworks

Two governance models are particularly relevant for reconceptualizing linguistic data portals: data commons and data trusts. Neither is a silver bullet. Both are imperfect. Precisely for that reason, they are useful — they provide concrete institutional logics for addressing the governance gaps the dominant narrative leaves unaddressed.

Elinor Ostrom's work on commons governance remains one of the most robust frameworks for thinking about shared resources (Ostrom 1990, 2015). On the surface, linguistic data appears non-rivalrous — multiple actors can use the same corpus without depleting it. But this apparent abundance is deceptive. What is at stake is not depletion, but enclosure: once linguistic resources are absorbed into proprietary AI pipelines, they may remain technically accessible while becoming functionally unavailable to the communities and institutions that produced them. Commons governance reframes this problem: instead of asking how data can be made maximally open, it asks who participates in setting the rules of access

and use. Ostrom's design principles — clearly defined boundaries, collective decision-making, monitoring, and conflict-resolution mechanisms — translate into concrete governance questions for linguistic data portals: who decides which actors can access the data? For what purposes? Under what conditions?

Where commons frameworks emphasize participation and collective rule-making, data trusts introduce a different logic: fiduciary responsibility (Delacroix/Lawrence 2019). In a data trust, trustees are legally and institutionally obligated to act in the interests of designated beneficiaries. For linguistic data, the trust model addresses a recurring governance gap: individual consent mechanisms are poorly suited to collective harms. A single speaker's consent cannot meaningfully authorize uses that affect an entire linguistic community, nor can it anticipate downstream applications of AI systems trained on aggregated language data. Data trusts respond to this mismatch by creating a governance layer that persists over time, rather than dissolving after initial data collection.

Hybrid models that combine commons and trust-based elements are the most promising arrangements. Participation without enforceability is fragile. Fiduciary oversight without participation is brittle. An instructive example is provided by federated research infrastructures such as CLARIN ERIC, which distribute linguistic data governance across national and institutional nodes while maintaining shared standards for access, interoperability, and ethical oversight. Although not formally structured as a data trust, CLARIN incorporates trust-like features — graduated access regimes, institutional accountability, and ethical review processes that constrain unrestricted reuse — within a participatory governance structure. This kind of hybrid arrangement, where commons-based mechanisms define shared values and access rules while trust-based mechanisms ensure those rules are upheld in interactions with powerful commercial actors, provides a concrete model for what governance-oriented linguistic data infrastructure can look like.

4.3 Indigenous Data Sovereignty and the CARE Principles

Any alternative framing of linguistic data risk that fails to engage with Indigenous data sovereignty is incomplete. Questions of data access, reuse, and control are inseparable from histories of dispossession, extractive

research practices, and the long-standing marginalization of Indigenous knowledge systems (Te Mana Raraunga 2018).

The CARE Principles — Collective Benefit, Authority to Control, Responsibility, and Ethics — were developed precisely to address the shortcomings of standard governance frameworks (Carroll et al. 2020). Authority to Control rejects the idea that consent is a one-time transaction, asserting ongoing governance authority over data collection, storage, access, and use. Responsibility extends obligations beyond data holders to data users. Ethics draws on Indigenous traditions that emphasize reciprocity, respect, and intergenerational responsibility. The risk of 'digital linguicide' — the quiet erosion of linguistic diversity through infrastructures that privilege dominant languages by default — is most acute for Indigenous and minority language communities. When public services, educational platforms, and AI systems function primarily in hegemonic languages, pressures toward language shift intensify. Governance failures here have cultural consequences that are irreversible on human timescales. The dominant narrative, which frames this as a coverage problem, cannot perceive it as irreversible loss.

4.4 The Alternative Narrative, the European Fundamental Rights Framework, and the Obstacles to Its Circulation

The counter-narrative proposed here can be understood as a socially communicable pattern of meaning that links actors, events, and causalities differently from the dominant account. In this counter-narrative, the actors are not primarily engineers and data scientists but linguistic communities, governance institutions, and data trustees. The problem is not insufficient data but unaccountable power over data. The causal mechanism is not scale but institutional design. The solution is not more data but better governance — governance anchored in rights, not merely in efficiency.

This counter-narrative finds important support in the European fundamental rights framework. Article 8 of the Charter of Fundamental Rights of the European Union enshrines the protection of personal data as an autonomous fundamental right, distinct from — and complementary to — the right to privacy under Article 7. The Court of Justice of the European Union has consistently interpreted this right substantively: in Schrems I and II, the CJEU established that data transfers cannot be authorized where the receiving legal framework fails to guarantee a level of protection essen-

tially equivalent to that ensured within the EU, and that technical or commercial arrangements cannot substitute for legal and institutional guarantees. Applied to linguistic data, this jurisprudential logic is directly relevant: data trusts and commons-based governance mechanisms are not merely efficiency innovations — they are institutional responses to a fundamental rights requirement. The GDPR's accountability principle (Article 5(2)), its purpose limitation requirement (Article 5(1)(b)), and its provisions on data protection impact assessments for high-risk processing (Article 35) all demand institutional structures capable of ensuring that data processing serves defined legitimate purposes and does not expose communities to disproportionate risks. The 'scale-and-fix' narrative, by foregrounding technical coverage while backgrounding institutional accountability, systematically undermines compliance with these requirements. Anchoring the alternative counter-narrative in the European fundamental rights framework transforms it from a normative aspiration into a legal obligation — one that data protection authorities, national supervisory bodies, and courts are positioned to enforce.

The effectiveness of this counter-narrative depends, as Müller-Funk (2002) and Amlinger (2023) emphasize, on its resonance with existing interpretive frameworks. The language of environmental stewardship — sustainability, biodiversity, ecological resilience — resonates broadly. The language of fiduciary obligation and collective rights resonates with legal and governance communities. The language of epistemic justice resonates with democratic theory and the sociology of knowledge. Mobilizing these resonances requires not just conceptual clarity but deliberate communicative strategy: identifying the audiences whose interpretive frameworks are closest to the alternative framing, and developing narratives pitched at the level of their existing understanding. The German data protection research community, in particular, brings a constitutional tradition that treats data protection as a fundamental right — a tradition that provides unusually fertile ground for narratives that foreground rights, accountability, and collective governance over technical adequacy. Yet significant obstacles to this counter-narrative's circulation must be honestly acknowledged. The first is institutional: the actors best positioned to produce and disseminate alternative narratives — Indigenous data sovereignty movements, civil society organizations, linguistically marginalized communities — are precisely those with the least access to the venues where dominant narratives are reproduced. International AI governance fora, technical standard-setting bodies, and major research funding agencies are structurally oriented

toward actors with established institutional presence and technical credibility. Alternative framings that challenge the solutionist consensus are systematically disadvantaged in these spaces, not because they lack merit, but because they resonate with interpretive frameworks — communitarian, rights-based, sovereignty-oriented — that are underrepresented in the dominant institutional culture of AI governance.

The second obstacle is conceptual: the counter-narrative requires its audience to perform a genuinely difficult cognitive shift — from perceiving linguistic data as a neutral resource to perceiving it as a politically contested commons embedded in power relations. This shift runs against deeply entrenched Deutungsmuster in technical communities, where data is understood as raw material and scale as progress. Overcoming these interpretive barriers requires sustained narrative work — not single publications, however well-argued, but iterative engagement across multiple venues, audiences, and media. The counter-narrative must be developed in forms accessible to data protection practitioners, communicated in language resonant with fundamental rights jurisprudence, and translated into policy proposals concrete enough to travel through existing regulatory channels. This is a long-term communicative project as much as an analytical one.

5. Conclusion: Governing Language Means Governing Meaning

This article has argued that the dominant discourse on linguistic diversity in artificial intelligence is best understood as a risk narrative — a socially stabilized pattern of meaning that renders certain risks visible and others invisible, and that actively shapes what counts as a problem, a solution, and a legitimate actor. The 'scale-and-fix' narrative misdiagnoses linguistic inequality in AI as a technical coverage problem, forecloses governance questions about control and accountability, and performs a systematic substitution of technical reassurances for institutional safeguards. A Bacchi-informed analysis of this narrative reveals what it leaves unproblematic: the extractive institutional arrangements that determine who controls linguistic data and for whose benefit, the epistemic consequences of homogenization, and the surveillance risks that technical fixes structurally cannot address.

Analysing this narrative through the lens of interpretive frameworks and socially shared patterns of meaning — rather than treating it as simply 'wrong' — illuminates why it persists. It is institutionally embedded. It resonates with widely shared solutionist assumptions. It travels efficiently

through policy documentation and technical discourse. And it provides plausible-sounding answers that relieve pressure for deeper institutional change. The counter-narrative developed here — reconceptualizing linguistic data portals as data trusts and data commons, governed through CARE principles and commons-based participation, anchored in the European fundamental rights framework — is not merely a governance proposal. It is an alternative sense-making structure: one that renders visible the structural and political dimensions of linguistic inequality, and opens space for accountability, collective rights, and democratic agency.

A final observation about the stakes of narrative choice. Governance narratives are not neutral descriptions. They determine what gets built, what gets funded, and what gets overlooked. If the dominant narrative of linguistic diversity in AI continues to frame the problem as one of coverage rather than governance, the erosion of linguistic plurality will not occur through dramatic acts of exclusion. It will occur quietly, through defaults, pipelines, and optimization objectives that appear reasonable in isolation. By the time the loss becomes visible, it may be difficult to reverse. The choice of narrative is, in this sense, already a governance decision — and, within the European fundamental rights framework, already a legal responsibility. That choice is being made now, largely by default. This article argues that it should be made deliberately, with full awareness of what the dominant story hides and whose rights it fails to protect.

References

Amlinger, C. (2023). Narrative. In: K. Dörr and C. Amlinger (Eds.), Populismus und Narrative. Springer VS.

Arnold, R. (1985). Deutungsmuster: Zu den Bedeutungselementen sowie den theoretischen und methodologischen Bezügen eines Begriffs. Zeitschrift für Pädagogik, 31(6), 893–912.

Arnold, R. (1991). Betriebliche Weiterbildung. Klinkhardt.

Bacchi, C. (2009). Analysing Policy: What's the Problem Represented to Be? Pearson Education.

Bacchi, C. (2012). Why Study Problematizations? Making Politics Visible. Open Journal of Political Science, 2(1), 1–8. https://doi.org/10.4236/ojps.2012.21001

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Polity Press.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. ACM Transactions on Information and System Security, 15(3), 1–22.

Carroll, S. R., et al. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal, 19(1), 43.

CLARIN ERIC. (n.d.). Common Language Resources and Technology Infrastructure. https://www.clarin.eu

Couldry, N., and Mejias, U. A. (2019). The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford University Press.

Court of Justice of the European Union. (2015). Schrems I. Case C-362/14, Maximillian Schrems v Data Protection Commissioner. ECLI:EU:C:2015:650.

Court of Justice of the European Union. (2020). Schrems II. Case C-311/18, Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems. ECLI:EU:C:2020:559.

Delacroix, S., and Lawrence, N. D. (2019). Bottom-up Data Trusts: Disturbing the 'One Size Fits All' Approach to Data Governance. International Data Privacy Law, 9(4), 236–252.

European Union. (2000). Charter of Fundamental Rights of the European Union. OJ C 364/01.

European Union. (2016). General Data Protection Regulation (GDPR). Regulation (EU) 2016/679 of the European Parliament and of the Council. OJ L 119/1.

Floridi, L. (2013). The Ethics of Information. Oxford University Press.

Floridi, L. (2014). The Fourth Revolution: How the Infosphere is Reshaping Human Reality. Oxford University Press.

Fricker, M. (2007). Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press.

Joshi, P., et al. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of ACL 2020, 6282–6293.

Kreutzer, J., et al. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Transactions of the Association for Computational Linguistics, 10, 50–72.

Küsters, I. (2022). Narrative Interviews. Grundlagen und Anwendungen (3rd ed.). Springer VS.

Kwet, M. (2019). Digital Colonialism: US Empire and the New Imperialism in the Global South. Race & Class, 60(4), 3–26.

Labov, W., and Waletzky, J. (1973). Erzählanalyse: Mündliche Versionen persönlicher Erfahrung. In J. Ihwe (Ed.), Literaturwissenschaft und Linguistik, Vol. 2, 78–126. Athenäum.

Longpre, S., et al. (2024). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. arXiv preprint arXiv:2408.04110.

Morozov, E., and Bria, F. (2018). Rethinking the Smart City: Democratizing Urban Technology. Rosa Luxemburg Stiftung.

Müller-Funk, W. (2002). Die Kultur und ihre Narrative. Springer.

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.

Ostrom, E. (2015). Governing the Commons (Canto Classics edition). Cambridge University Press.

Rosa, J., and Flores, N. (2017). Unsettling Race and Language: Toward a Raciolinguistic Perspective. Language in Society, 46(5), 621–647.

Siebert, H. (1996). Didaktisches Handeln in der Erwachsenenbildung: Didaktik aus konstruktivistischer Sicht. Luchterhand.

Te Mana Raraunga. (2018). Principles of Māori Data Sovereignty. https://www.temanaraunga.maori.nz/

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.

About the Author

Prof. Paola Cantarini Guerra

is a lawyer and university professor. PhD in Philosophy, in Philosophy of Law, and in Law. Researcher in AI governance, technology ethics, and digital constitutionalism. Coordinator of AI governance research projects at the University of São Paulo and visiting researcher at the Alexander von Humboldt Institute for Internet and Society (Berlin).