Polycentric control beyond compliance: A comparative qualitative study of six contemporary AI Governance models

Aníbal Astobiza Monasterio[1], Cristina Astier[2,3], and Txetxu Ausín[4]

[1] Departamento de Filosofía I, Universidad de Granada, Granada, Spain
[2] Departamento de Derecho Administrativo, Constitucional y Filosofía del Derecho, EHU, Leioa, Spain
[3] Departamento de Derecho, Universitat Pompeu Fabra, Barcelona, Spain
[4] Instituto de Filosofía, CSIC, Madrid, Spain
amastobiza@ugr.es, cristina.astier@upf.edu, mariacristina.astier@ehu.eus, txetxu.ausin@cchs.csic.es

Abstract. Polycentric governance is often presented as the normative answer to the concentration of power in artificial intelligence (AI), yet current debates rarely show how polycentric control can be identified and compared across concrete governance models. This paper addresses that gap through a qualitative comparative document analysis of six contemporary governance models published or updated between 2024 and 2026: the EU AI Act, the NIST Generative AI Profile, the United Nations AI governance initiatives, a power-sharing liberal roadmap, a public participation model for public administration, and a coordination transparency framework. The study codes each model across four dimensions: decisional plurality, citizen contestability, runtime oversight, and public participation. Scoring follows an anchored codebook with explicit 0–2 descriptors and decision rules, and the full scoring matrix and resulting ranking are reported. The results show that current approaches cluster into three incomplete families: compliance-centered models, enforceable but weak on democratic inclusion; participation-centered models, normatively attractive but weak on continuous technical intervention; and coordination-centered models, which illuminate distributed agency but remain thinly institutionalized. No single model integrates all four dimensions. The paper therefore proposes a Hybrid Polycentric Control model that combines legal enforceability, participatory channels, and runtime coordination transparency. The contribution is methodological and empirical: it operationalizes polycentric control as a measurable, multi-dimensional property of governance texts and uses the resulting instrument to clarify the operational demands of democratic control over AI.

Keywords: polycentric governance · AI governance · democratic control · contestability · runtime oversight · sociotechnical systems.

1 Introduction

Artificial intelligence (AI) governance has moved in a short period from abstract ethical principles to a denser institutional field: the EU AI Act, the NIST Generative AI

Profile, and the United Nations initiatives show that the policy debate is no longer limited to voluntary ethics statements [8,9,10]. Yet this institutional thickening has not resolved a central political problem: many current approaches still concentrate agenda setting, monitoring, or intervention capacity in a narrow set of governmental, corporate, or expert actors. Recent scholarship therefore describes a democratic deficit in AI governance [2] and calls for an AI governance model capable of protecting values embedded in a democratic society [12].

Polycentric governance is attractive in this context because it promises to distribute authority across multiple decision centers instead of relying on a single regulator, laboratory, or compliance office [1]. The difficulty is that the AI literature often invokes polycentricity as an aspiration rather than measuring it as an institutional property [13,14,15]: a governance arrangement may involve many actors while leaving meaningful control centralized, and mere multiplicity is not democratic control.

Within this framework, this paper asks when contemporary AI governance models produce substantive democratic control instead of symbolic inclusion. To answer that question, it conducts a qualitative comparative document analysis of six governance models published or updated between 2024 and 2026 and codes them across four dimensions: decisional plurality, citizen contestability, runtime oversight, and public participation, each scored on an anchored 0–2 scale defined in an explicit codebook, so that every point assignment can be traced back to quotable provisions of the source texts. The central finding is that current approaches cluster into three incomplete families, compliance-centered, participation-centered, and coordination-centered, none of which integrates all four dimensions. On that basis, the paper proposes a Hybrid Polycentric Control model that combines legal enforceability, participatory channels, and coordination transparency.

The contribution is therefore threefold: the paper turns polycentric control into a small comparative instrument, four dimensions, an anchored scoring rubric, and a replicable coding procedure supported by a semi-automated consistency check (Appendix A); it applies that instrument to six influential governance models, reporting the full scoring matrix, the per-dimension justifications, and the aggregate ranking; and it uses the observed pattern of complementary strengths and blind spots to motivate a hybrid design rule for future governance arrangements.

2 Background and analytical framework

The recent literature confirms both the urgency and the fragmentation of AI governance. Batool et al. show in a systematic review that the field spans legal, ethical, organizational, and technical approaches, but lacks a stable integrative vocabulary [3]. In this line, Coeckelbergh argues that AI governance remains vulnerable to a democratic deficit when citizens and affected publics are weakly represented in decisions that shape the design and deployment of AI systems [2].

Several recent contributions address pieces of the polycentric puzzle. In AI governance, a polycentric model involves different epistemic communities and multiple centers of decision-making power operating with a certain level of autonomy [13]. Allen et al. propose a roadmap for power-sharing liberalism that shifts the focus from technical

alignment to the distribution of decision rights [4]. Taylor et al. argue that public participation must be treated as a substantive component of public-sector AI governance rather than a symbolic afterthought [5]. Bohr shows that human control over AI systems increasingly depends on coordination transparency, that is, on visibility into how distributed technical and organizational actors interact over time [6]. Dahlgren et al. warn that Reinforcement Learning from Human Feedback (RLHF), the technique by which large models are tuned to human preference judgments, and adjacent alignment methods have sociotechnical limits and cannot substitute for institutional governance [7].

2.1 From Polycentric aspiration to measurable control

The analytical gap we identify is not the absence of normative concern with polycentric AI governance, but the absence of a common basis for comparison within a framework of substantive democratic control. Official frameworks tend to emphasize accountability, documentation, or risk management [8,9]. Participatory models emphasize legitimacy and inclusion [5]. Coordination-focused accounts emphasize the runtime problem of distributed agency [6]. These elements are usually discussed in isolation. Bridging this gap, the present study treats these elements as complementary dimensions of polycentric control and asks how they are combined, or left uncombined, across current governance models.

This is also where the study differs from existing accounts of the polycentric AI governance landscape [13,14,15], which map the field, identifying the actors, regimes, and epistemic communities among which authority is dispersed and diagnosing fragmentation or convergence at the level of the regime complex. What they do not provide is an instrument for asking, of any particular model, how much control it actually redistributes and to whom; our contribution is precisely that operationalization. By decomposing polycentric control into four scoreable dimensions, the framework distinguishes shallow polycentricity, in which many actors are present but few can decide, contest, or intervene, from effective polycentric control, in which those capacities are institutionally secured. The framework also foregrounds a dimension that mapping approaches leave under-specified: runtime oversight, the persistence of monitoring and intervention capacity after deployment.

3 Materials and methods

3.1 Corpus

The corpus contains six governance models that are highly visible, recent, and substantively different in design: the EU AI Act [8], the NIST Generative AI Profile [9], the United Nations AI governance initiatives [10], the power-sharing liberal roadmap proposed by Allen et al. [4], the public participation model discussed by Taylor et al. [5], and Bohr's coordination transparency framework [6]. The choice is purposeful rather than exhaustive: the aim is not to represent the entire governance literature, but to compare influential models, each capturing a distinct route toward control over AI systems.

Three inclusion criteria governed the selection: recency, every model was published or substantially updated between 2024 and 2026; institutional influence, each model is legally binding in a major jurisdiction, issued by a standard-setting body or by the United Nations, or prominently discussed in the recent literature; and maximal design variation, the six models deliberately span the compliance, participation, and coordination corners of the design space, following a most-different-systems logic. Models that differ this much in origin, legal status, and normative orientation make it unlikely that shared structural gaps are an artifact of sampling from a single tradition.

The empirical strategy is qualitative comparative document analysis: each source text has been analyzed for explicit provisions or institutional proposals concerning who may decide, who may intervene, who may contest decisions, and how oversight persists after deployment. The unit of analysis is the governance model as articulated in the source text, not the downstream performance of any organization that adopts it.

3.2 Coding scheme

The coding scheme uses four dimensions as a common ground for assessing whether the analyzed models produce substantive democratic control instead of symbolic inclusion. The first dimension, decisional plurality, asks whether authority is distributed across more than one meaningful center. The second, citizen contestability, asks whether affected persons can question, challenge, or redirect AI-mediated decisions. The third, runtime oversight, asks whether monitoring and intervention persist during use, not just before deployment. The fourth, public participation, asks whether lay publics or civil society actors are institutionally included rather than merely informed after the fact.

Each dimension is scored on a three-point scale, from 0 (absent or merely rhetorical) through 1 (partially specified) to 2 (explicit, actionable institutional design). The scoring remains interpretive, but it is not unconstrained: every point assignment follows the anchored rubric presented next and must be traceable to at least one quotable passage of the source text.

3.3 Scoring rubric

Table 1 presents the codebook: each cell states the anchored descriptor that a source text must satisfy to receive the corresponding score.

Three decision rules complete the codebook. First, only explicit provisions or institutional proposals count as evidence: aspirational language in preambles, visions, or recitals can never raise a score. Second, when a text oscillates between two anchors, the lower score is assigned unless the relevant provision is formulated as a mandatory requirement rather than a recommendation. Third, each score must be documented with at least one supporting excerpt, so that any reader can audit the assignment against the rubric.

3.4 Coding procedure and validation

The coding proceeded in three steps. First, the source texts were read in full and segmented into provisions relevant to any of the four dimensions. Second, each model

Table 1. Anchored codebook: descriptors for scores 0, 1, and 2 on each dimension.

Dimension	0 (absent/ rhetorical)	1 (partially specified)	2 (actionable design)
Decisional plurality	Authority rests with a single center; plurality only invoked	Several bodies named, but within one hierarchy or without autonomous mandates	Two or more centers hold autonomous mandates and explicit decision rights
Citizen contestability	No channel for affected persons to challenge AI-mediated decisions	Contestation mentioned, but addressee, object, or procedure unspecified	Explicit complaint or review mechanism with identified addressee and procedure
Runtime oversight	Oversight confined to design or pre-deployment conformity	Post-deployment monitoring mentioned, without intervention mechanism	Explicit monitoring plus intervention, interruption, or incident-response duties during use
Public participation	Publics informed after the fact, or not addressed	Consultation foreseen but episodic or advisory	Institutionalized participation with defined role in decisions

was scored on each dimension according to the codebook, with the supporting excerpts recorded alongside every score; disagreements among the authors were adjudicated by discussion against the anchored descriptors. Third, the semi-automated coder described in Appendix A was run on the same corpus as an independent consistency check, flagging cells where the automatically retrieved evidence appears weaker or stronger than the human score. The complete coding materials, scripts, prototypes, and per-model outputs with supporting excerpts are archived in the public repository referenced in the Code availability statement.

We remain candid about what this procedure can establish. On a purposive sample of six documents, formal inter-coder agreement statistics would carry little evidential weight, and we do not report them; what the procedure secures is traceability. Every cell of the matrix in Section 4 can be checked against the codebook and archived excerpts, and the automated check guards against treating similar documents inconsistently, the most common failure of interpretive coding.

4 Results

The comparative reading yields a clear pattern: compliance frameworks are strongest on legal duties, documentation, and supervision, participatory frameworks on inclusion and legitimacy, and coordination-centered approaches on the runtime problem of distributed agency. What is missing is a single model that unifies these strengths into one coherent architecture of substantive democratic control. The finding is reported in two complementary forms: the scoring matrix with its ranking (Table 2) and a qualitative summary of dominant strengths and blind spots (Table 3).

4.1 Scores and ranking

Table 2 presents the scores assigned under the codebook of Section 3.3; the following paragraphs condense the justification for each assignment.

Table 2. Scoring matrix: scores (0–2) of the six governance models on the four dimensions, with unweighted totals. DP = decisional plurality; CC = citizen contestability; RO = runtime oversight; PP = public participation.

Model	DP	CC	RO	PP	Total
Allen et al. [4]	2	1	1	2	6
EU AI Act [8]	1	1	2	1	5
Taylor et al. [5]	1	2	0	2	5
UN AI initiatives [10]	2	0	0	2	4
NIST GenAI Profile [9]	0	0	2	1	3
Bohr [6]	1	0	2	0	3

The compliance-centered models score highest on runtime oversight and lowest on the democratic dimensions. The EU AI Act receives 2 on runtime oversight for its explicit post-market monitoring, serious-incident reporting, and human-oversight duties for high-risk systems; 1 on decisional plurality, because the European AI Office, national market surveillance authorities, and notified bodies are plural but operate within a single regulatory hierarchy; 1 on citizen contestability, because complaint and explanation rights run through supervisory authorities with substantial friction; and 1 on public participation, because stakeholder input remains advisory and episodic [8]. The NIST profile receives 2 on runtime oversight for its monitoring, incident-response, and decommissioning guidance and 1 on public participation for its recommendation of external stakeholder feedback, but 0 on the remaining dimensions: as a voluntary framework, it neither distributes authority beyond the adopting organization nor creates any channel for affected persons [9].

The participation-centered models show the mirror-image profile. The UN initiatives receive 2 on decisional plurality, because the Independent International Scientific Panel and the Global Dialogue constitute distinct centers with their own mandates alongside member states, and 2 on public participation for their institutionalized multistakeholder design, but 0 on citizen contestability and runtime oversight, since no channel of challenge or intervention capacity is specified [10]. Taylor et al. receive 2 on citizen contestability and public participation, since maximum feasible participation gives affected communities a defined role in administrative decisions and an avenue to challenge them; 1 on decisional plurality for the partially specified community bodies alongside public agencies; and 0 on runtime oversight, which the proposal does not address [5]. Allen et al. receive 2 on decisional plurality, because redistributing decision rights is the explicit organizing principle of power-sharing liberalism, and 2 on public participation for the roadmap's democratic innovations; contestation and runtime arrangements are sketched rather than designed, yielding 1 on both remaining dimensions [4].

Bohr's coordination transparency framework receives 2 on runtime oversight, since continuous, interruptible visibility into distributed interactions is precisely what the framework demands, and 1 on decisional plurality for its recognition of distributed agency. It receives 0 on citizen contestability and public participation: the account specifies who must see the coordination structure, but not who may challenge it or how publics enter the arrangement [6].

Summing the four dimensions yields the aggregate ranking: Allen et al. rank first with 6 points, the EU AI Act and Taylor et al. share second position with 5, the UN initiatives follow with 4, and the NIST profile and Bohr close with 3. The total is an unweighted, ordinal transparency device, not a cardinal measure of governance quality: a different weighting would reorder the middle of the table without affecting the extremes. The columns, however, are more informative than the rows: no model in the corpus scores 2 on more than two dimensions, and every model scores 0 or 1 on at least one. That column-wise incompleteness is the central empirical result of the study.

4.2 Strengths and blind spots

Stated qualitatively, Table 3 summarizes the dominant strength and blind spot of each model: the matrix records the intensity of each capacity, the table its kind.

Table 3. Strengths and blind spots of the six governance models in the corpus.

Model	Dominant strength	Dominant blind spot
EU AI Act	Enforceable lifecycle duties and supervision	Limited direct citizen contestability
NIST AI RMF	Operational risk language and monitoring logic	Voluntary and organization-centered
UN AI initiatives	Global multi-stakeholder legitimacy	Weak enforcement and runtime intervention
Taylor et al.	Strongest participation design	Thin technical lifecycle governance
Allen et al.	Most explicit redistribution of power	Limited implementation detail
Bohr	Strongest runtime coordination requirement	Weak democratic institutionalization

4.3 Three recurrent deficiencies

The matrix makes three recurrent deficiencies visible as column patterns. The first deficiency is compliance without democratization. The EU AI Act and the NIST profile offer vocabulary, procedures, and documentation that can discipline organizations, but

they do not guarantee strong public participation or low-friction contestability. The second deficiency is participation without runtime control. Taylor et al. and the UN initiatives show how governance can become more inclusive, but they provide less detail on how intervention rights are preserved once systems continue learning, interacting, or coordinating after deployment; together they score a single point on runtime oversight. The third deficiency is coordination without institutional embedding. Bohr identifies the need for coordination transparency, but that insight remains weak unless linked to legal duties, escalation pathways, and public-facing oversight.

Taken together, these results suggest that the relevant contrast is not between centralized and polycentric governance in the abstract, but between shallow polycentricity, where many actors comment on AI, and effective polycentric control, where authority, contestation, and intervention are genuinely redistributed. That is the point at which polycentric governance becomes more than a legitimizing metaphor.

5 Discussion

The study supports a stronger claim than the standard multi-stakeholder formula. Polycentric governance improves AI democratic control only when it links at least three layers that are currently fragmented in the literature and in policy. First, a compliance layer creates enforceable duties, documentation, and supervisory authority. Second, a participatory layer creates legitimate channels for public input, challenge, and revision. Third, a runtime coordination layer makes distributed technical and organizational interactions visible and interruptible.

5.1 A Hybrid Polycentric Control model

The Hybrid Polycentric Control model proposed here is therefore not a new institution but a design rule: robust AI governance requires the joint presence of legal enforceability, public contestability, and coordination transparency. In practical terms, risk management documents should be linked to public-facing summaries, affected parties should have intelligible review and escalation pathways, and oversight should extend beyond pre-deployment conformity to the runtime interactions that shape outcomes in sociotechnical systems.

This formulation is especially relevant for human-machine collaboration systems, where the locus of agency is distributed across users, models, interfaces, organizations, and institutional rules. In such environments human oversight cannot be reduced to a single person in the loop; it depends on whether multiple actors can see, question, and interrupt the coordination structure itself. Coordination transparency should therefore be treated as a democratic design requirement, rather than a purely technical feature [6].

An arrangement satisfies the hybrid requirement when it would score 2 on all four dimensions of the codebook, and Table 2 shows how the existing models could supply the missing pieces to one another: the EU AI Act already contributes the enforceable runtime layer, the participation-centered models specify exactly what it lacks, and the coordination-centered account provides the visibility conditions under which either

layer can operate. The hybrid model is thus not a speculative construct but a recombination of provisions that already exist, separately, in the current governance generation. Finally, the matrix clarifies the division of labor between technical alignment and institutional governance: alignment methods such as RLHF operate on the internal dispositions of a model, but they do not distribute decision rights, create channels of contestation, or render organizational coordination visible [7]; they are complements to, never substitutes for, institutional architecture.

5.2 Limitations

The study has obvious limits. The corpus is small and purposefully selected: six models chosen for maximal design variation support analytic, not statistical, generalization. The coding remains interpretive: the codebook constrains judgment and makes every score auditable, but it does not eliminate judgment, and readers may contest individual cells of the matrix. The unweighted aggregation treats the four dimensions as equally important, an explicit normative choice. Finally, the paper analyzes governance models as documents rather than tracing their adoption in real organizations. These limits matter, but they do not undermine the central contribution: the comparison suffices to show that current governance debates repeatedly separate inclusion, enforceability, and runtime control, and the instrument that produces this finding is fully specified, archived, and reusable on larger corpora.

6 Conclusion

Polycentric governance should not be treated as a synonym for stakeholder multiplicity. The comparative analysis presented here shows that substantive democratic control over AI depends on a more demanding institutional combination, namely, multiple decision centers, credible channels for contestation, and continuous visibility into distributed coordination. Current governance models each illuminate one part of that problem, but none solves it alone: no model in the corpus reaches the top score on more than two of the four dimensions. The paper contributes a compact instrument for making that diagnosis, an anchored codebook, a documented scoring procedure, and a semi-automated consistency check, that others can reuse. Future AI governance should be designed not only to assign responsibility, but also to redistribute control in ways that remain visible, challengeable, and interruptible over time. The overall objective of any AI governance model should be to operationalize the values embedded in democratic societies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Code availability. The notebook and scripts used for the semi-automated coding procedure are archived in a public repository with persistent identifier: https://doi.org/10.5281/zenodo.19196034

For executable access, a Google Colab mirror is available at: https://colab.research.google.com/github/anibalmastobiza/polycentric-control-lnai-materials/blob/main/colab_rbf_governance_coder.ipynb

References

1. Ostrom, E.: Beyond markets and states: polycentric governance of complex economic systems. American Economic Review 100(3), 641-672 (2010)
2. Coeckelbergh M. Artificial intelligence, the common good, and the democratic deficit in AI governance. AI and Ethics. 2025 5(2):1491-7. https://doi.org/10.1007/s43681-024-00492-9
3. Batool A, Zowghi D, Bano M. AI governance: a systematic literature review. AI and Ethics. 2025 Jun;5(3):3265-79. https://doi.org/10.1007/s43681-024-00653-w
4. Allen D, Hubbard S, Lim W, Stanger A, Wagman S, Zalesne K, Omoakhalen O. A roadmap for governing AI: technology governance and power-sharing liberalism. AI and Ethics. 2025 5(3):3355-77. https://doi.org/10.1007/s43681-024-00635-y
5. Taylor RR, Murphy JW, Hoston WT, Senkaiahliyan S. Democratizing AI in public administration: improving equity through maximum feasible participation. AI & society. 2025 40(5):3653-62. https://doi.org/10.1007/s00146-024-02120-w
6. Bohr J. Coordination transparency: governing distributed agency in AI systems. AI & SOCIETY. 2026 15:1-1. https://doi.org/10.1007/s00146-026-02853-w
7. Dahlgren Lindström A, Methnani L, Krause L, Ericson P, de Rituerto de Troya ÍM, Coelho Mollo D, Dobbe R. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. Ethics and Information Technology. 2025 27(2):28. https://doi.org/10.1007/s10676-025-09837-2
8. European Union: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union (2024). https://eur-lex.europa.eu/legal-content/EN-FR/TXT/?from=EN&uri=CELEX%3A32024R1689
9. NIST: Artificial Intelligence Risk Management Framework: Generative AI Profile. National Institute of Standards and Technology, Gaithersburg (2024). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
10. United Nations: General Assembly establishes Independent International Scientific Panel on AI and Global Dialogue on AI Governance. United Nations, New York (2025). https://www.un.org/global-digital-compact/en/ai
11. Broomhead, D.S., Lowe, D.: Multivariable functional interpolation and adaptive networks. Complex Systems 2(3), 321-355 (1988)
12. Innerarity, D. Una Teoría Crítica de la Inteligencia Artificial. Barcelona: Galaxia Guttenberg, 2025.
13. Schmitt, L. Mapping global AI governance: a nascent regime in a fragmented landscape. AI Ethics 2, 303–314 (2022). https://doi.org/10.1007/s43681-021-00083-y
14. Woods, D. From fragmentation to polycentricity: a comparative synthesis on AI governance and global policy coordination. Glob. Public Policy Gov. 5, 87–92 (2025). https://doi.org/10.1007/s43508-025-00118-8
15. Xue JH. Polycentric Theory Diffusion and AI Governance. In: Global Digital Data Governance 2024 (pp. 223-237). Routledge.

A Semi-automated coding protocol

The semi-automated coder used as a consistency check in Sect. 3.4 is a radial basis function (RBF) prototype scorer: a fixed, fully inspectable similarity procedure that borrows the RBF kernel from the classical RBF literature [11]. It is not a trained RBF network: fitting parameters to labeled data would not be credible on six documents, and the coder contains no learned parameters.

Each document is segmented into paragraphs or, where these are long, sentences, each represented as a TF-IDF vector x over the corpus vocabulary. For each dimension we hand-wrote prototype passages expressing what a maximally explicit provision would say, represented in the same space as vectors p. For every segment–prototype pair the script computes the RBF-kernel similarity $s(x,p) = \exp(-\gamma\|x-p\|^2)$, with γ a fixed width parameter. For each document and dimension the coder aggregates the maximum similarity, the coverage (segments above a moderate threshold), and the count of dimension-specific actor and action cues. Fixed decision rules map these quantities onto the 0–2 scale of Table 1: 2 requires high maximal similarity plus at least one segment combining cue evidence with actionable phrasing; 1 requires moderate similarity or cue evidence without actionable phrasing; 0 is assigned otherwise. All thresholds are stated in the released code and were fixed in advance.

The coder's outputs were compared cell by cell against the adjudicated human scores of Table 2; where the first pass diverged, the cell was re-examined against the codebook and the excerpts, and the human score was retained only when the divergence stemmed from a limitation of the surface representation, typically cross-references invisible to TF-IDF similarity. The complete outputs, with the evidence passages for every cell, are archived alongside the scripts. The automation is thus a reproducibility instrument, not a claim of machine-learned validity.