Measuring What Matters: The AI Pluralism Index

Rashid Mushkani
Right to AI
Université de Montréal
Mila – Québec AI Institute
rashidmushkani@gmail.com

Abstract

Artificial intelligence systems increasingly mediate knowledge, communication, and decision making. Development and governance remain concentrated within a small set of firms and states, raising concerns that technologies may encode narrow interests and limit public agency. Capability benchmarks for language, vision, and coding are common, yet public, auditable measures of pluralistic governance are rare. We define *AI pluralism* as the degree to which affected stakeholders can shape objectives, data practices, safeguards, and deployment. We present the AI Pluralism Index (AIPI), a transparent, evidence-based instrument that evaluates producers and system families across four pillars: *participatory governance*, *inclusivity & diversity*, *transparency*, and *accountability*. AIPI codes verifiable practices from public artifacts and independent evaluations, explicitly handling *Unknown* evidence to report both lower-bound ("evidence") and known-only scores with coverage. We formalize the measurement model; implement a reproducible pipeline that integrates structured web and repository analysis, external assessments, and expert interviews; and assess reliability with inter-rater agreement, coverage reporting, cross-index correlations, and sensitivity analysis. The protocol, codebook, scoring scripts, and evidence graph are maintained openly with versioned releases and a public adjudication process. We report pilot provider results and situate AIPI relative to adjacent transparency, safety, and governance frameworks. The index aims to steer incentives toward pluralistic practice and to equip policymakers, procurers, and the public with comparable evidence.

Index: https://aipluralism.wiki/
Code: https://github.com/rsdmu/aipi-pluralism-index

1 Introduction

Artificial intelligence is now infrastructural in search, education, media, health, finance, and public administration. Decisions about data curation, model objectives, deployment defaults, safety mitigations, and redress are largely made by a small number of labs and government programs. International guidance warns that such concentration undermines democratic legitimacy and risks sustained exclusion of those most affected [42, 9, 3, 8]. Users face a second challenge: for many everyday tasks, top systems often produce broadly similar outputs; headline capability differences are therefore less informative for responsible selection [35, 6, 7, 38]. When functionality converges, the governance properties of systems become a salient differentiator.

Existing evaluations compare model capabilities and user preferences [35, 7, 39]. Other efforts synthesize ethics and disclosure guidance [23, 18, 30, 5, 22] and safety governance [2, 15, 11, 21]. Regulatory frameworks articulate process expectations—risk management, incident handling, documentation, accessibility—but do not themselves yield comparable scores across providers [40, 12].

Article presented at the Second IASEAI Conference (IASEAI'26).

There is no comprehensive, public, auditable instrument that centers pluralism—whether affected stakeholders can influence objectives, safeguards, and deployment—and measures participation, inclusion, transparency, and accountability in verifiable ways. This paper addresses that gap.

Contributions. This paper introduces the AI Pluralism Index (AIPI), a community-governed protocol that: (i) formalizes an indicator hierarchy operationalizing pluralism across four pillars grounded in international norms and empirical work [23, 1, 4, 27, 16, 17, 42, 32, 14, 43]; (ii) implements a transparent evidence pipeline and scoring with two principled treatments of *Unknown*; (iii) evaluates reliability and construct validity via inter-rater agreement, cross-index correlations, and sensitivity analysis; and (iv) publishes system- and provider-level results with evidence provenance through a versioned, open governance process aligned with standards for incident reporting and vulnerability disclosure [19, 20, 37].

2 Related work

Capability benchmarks such as Chatbot Arena and HELM provide large-scale comparative results for tasks under controlled protocols [35, 7, 39]. These efforts inform model selection for performance but are largely orthogonal to governance. Transparency syntheses assess disclosure practices and ethics guidance [23, 18, 5, 22]. Safety indices and governance assessments focus on risk management commitments and safeguards [2, 15]. Regulatory instruments set expectations for trustworthy AI without creating measurement tools: the EU AI Act defines risk-based obligations, documentation, and oversight [12]. The NIST AI Risk Management Framework provides organizational process scaffolding [40, 21]. AIPI complements these efforts by offering an auditable, evidence-driven index focused on pluralism: participation, inclusion, transparency, and accountability practices that shape how systems are built and governed.

3 Concept and scope

We define **AI pluralism** as a property of socio-technical development and deployment processes whereby affected stakeholders can meaningfully shape objectives, constraints, and outcomes. Pluralism concerns both governance and production. The index is anchored in four equal-weighted pillars (Fig. 1): participatory governance, inclusivity and diversity, transparency, and accountability. These pillars align with traditions in participation studies and human–computer interaction—Arnstein's ladder and deliberative traditions emphasize degrees of influence and institutional design [1, 4, 14, 29]; Value Sensitive Design operationalizes stakeholder engagement and value articulation in technology development [41, 13]. Inclusivity requires attention to representation in teams, language access, and accessibility [17, 43]. Transparency and accountability cover documentation, incident reporting, audits, redress, and vulnerability handling [27, 16, 25, 19, 20]. The unit of analysis is the *producer* (company, lab, consortium, or public agency), with optional system-family granularity when practices differ across offerings. The scope is global and normative, aligning with recognized principles rather than a single jurisdiction. Because AIPI is intended as a global index, participatory governance indicators distinguish between geographically bounded engagement (e.g., only in a producer's home jurisdiction) and engagement that is demonstrably multi-region. We do not apply separate region or population weights in aggregation; instead, geographic scope is captured within relevant indicator rubrics and reported as metadata in the evidence graph.

4 Design choices: indicators, observability, and consequences

AIPI credits only practices that leave public, auditable traces. Indicators must be *observable* through public artifacts or external evaluations, *verifiable* via durable references and timestamps (with archives when permitted), and *consequential* in the sense that they allocate decision rights, impose process requirements, or document changes. This approach reduces coder discretion, enables third-party audit, and aligns scoring with practices that affected stakeholders can inspect and use. More broadly, AIPI follows established methodology for constructing composite indicators, including explicit conceptual framing, transparent normalization and weighting choices, and sensitivity analysis [33].

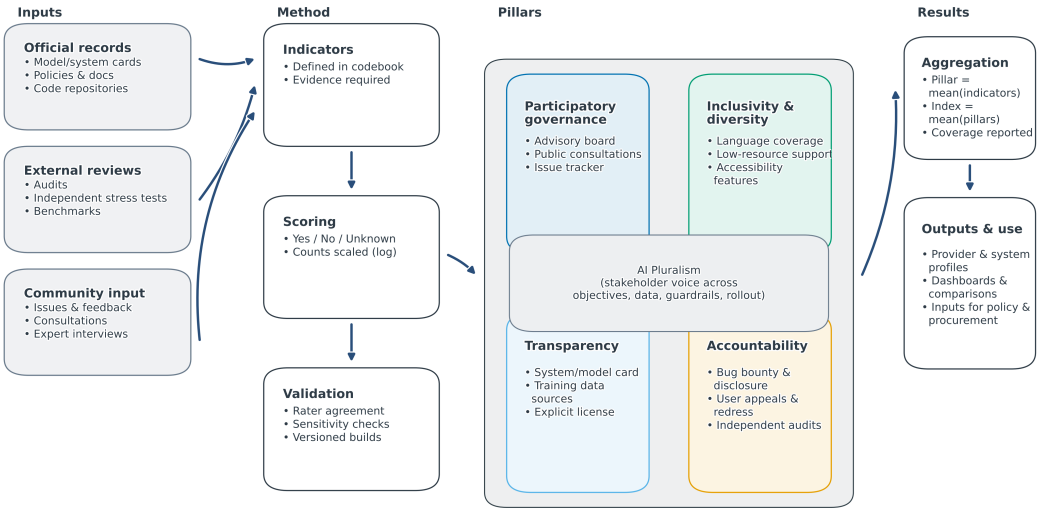

Figure 1: Conceptual framework of the AI Pluralism Index (AIPI). Verifiable evidence is coded into indicators, aggregated into four equal-weighted pillars (Participatory governance, Inclusivity & diversity, Transparency, and Accountability), and averaged to produce AIPI. Open governance, versioned releases, and validation provide auditability.

Observability requires a primary artifact rather than a claim about one. Acceptable artifacts include policies, model or system cards, release notes, governance records (e.g., minutes, charters, response summaries), audit reports, and entries in recognized registries. To receive full credit, an artifact must identify the decision locus (who decides and on what scope), describe the process or scope (e.g., advisory remit, audit coverage, data governance components), and record outcomes or dispositions (e.g., decisions, remediation, linked change logs). A model card that specifies purpose, training data provenance, intended uses, limitations, subgroup performance, and update history is observable evidence [27, 16]. An independent audit that publishes scope, method, findings, and remediation is observable evidence [25, 34]. Marketing pages or press releases without scope, process, or outcomes are not.

Verifiability requires that evidence be durable and comparable across releases. Each artifact must have a persistent URL and a clear date; where possible we store or reference an archived copy. Versioned artifacts are coded by version. Materials behind paywalls, NDAs, or logins do not qualify as primary evidence because affected stakeholders cannot inspect them. When a standards claim is made (e.g., "WCAG 2.2 AA"), the artifact should state scope (surfaces, products, versions) and date; claims without scope or date receive partial credit under the project's generic ordinal rubric [17, 43].

Consequence is assessed by whether a practice changes procedures, authority, or obligations. In participatory governance, credit attaches to mechanisms that record input and document dispositions, such as public consultations with response summaries, advisory bodies with published remits and minutes, and external community red teaming with published scope and outcomes [1, 14]. Generic feedback channels without dispositions do not receive credit.

For inclusivity and diversity, we count language support only when multiple user-facing components are present for the flagship surface. A language is counted if at least three of the following are documented: localized user interface, localized help or FAQ, localized policies and safety disclosures, and a localized redress or appeals channel. Auto-translation or preview-only modes are not counted. Accessibility claims are assessed with stated scope and date; self-declarations without scope receive partial credit; independent verification receives full credit [17, 43]. Partnerships outside high-income markets count when they involve co-production (e.g., co-authored resources, shared maintenance) rather than one-off donations. Statements of commitment without accompanying statistics, methods, or changes do not receive credit. For multilingual evaluation and collaboration in lower-resourced settings, we draw on evidence from participatory language work [31].

Transparency indicators credit documentation that is decision-relevant for users, procurers, and auditors. Model and system cards are coded when they cover purpose, data provenance, intended uses, limitations, subgroup results, and update history [27, 16]. Release notes that link governance or safeguard changes to dates and versions are counted. Narrative blog posts without an associated artifact are not counted.

Accountability indicators distinguish policy statements from implemented processes. Vulnerability disclosure is credited when inbound channels, scope, and safe-harbor language are documented; vulnerability handling is credited when roles, triage, and timelines are documented, with references to recognized practice [25, 26]. Independent audits with public findings and remediation timelines are credited above self-attestations. Incident postmortems are credited when they supply technical and procedural details; contributions to public learning repositories and impact assessment practices are counted as additional evidence [36, 37].

Where mature standards exist, coding is anchored to them to reduce ambiguity across providers and releases. For accessibility we assess the presence of concrete conformance claims and independent verification [17]; for security disclosure and handling we look for documented processes [25, 26]. Standards references without scope or date do not suffice; conformance statements with scope and date receive partial credit; verification (e.g., audit or test report) receives full credit [43, 19, 20].

Counting indicators (e.g., number of supported languages, consultations, or disclosed incidents) follow inclusion rules designed to limit inflation. Counts are mapped using the tempered logarithmic transform in Eq. (1) to cap marginal influence and reduce sensitivity to outliers (Sec. 5). Disclosed incidents are not penalized; when paired with postmortems they are treated as accountability evidence.

Boundary cases follow consistent rules. Primary artifacts take precedence over secondary reporting; without a primary artifact, no credit is assigned. Provider-level artifacts apply to all systems only when scope is explicit; otherwise credit is limited to the named systems. Artifacts predating the evidence window are coded but may receive partial credit for staleness. Gated materials do not qualify as observable. When public claims conflict, we code to the more conservative interpretation and record the discrepancy in adjudication logs.

Certain items are excluded by design: principle statements without mechanisms; advisory bodies without a charter, roster, cadence, and outputs; redress channels that are indistinguishable from generic support (no service levels, no appeal paths); language claims based only on translation without localized policies or redress; and private assurances made to the project team.

The scoring implications are direct. In the evidence model, missing public artifacts are treated as zero, which assigns credit only when documentation exists. The known-only model averages over observed indicators. Coverage is reported to separate documentation gaps from practice and to signal where rankings may be sensitive to missing evidence (Sec. 5).

5 Measurement model

AIPI aggregates evidence-coded indicators into pillar and overall scores with explicit treatment of *Unknown*. Figure 2 provides a visual overview of the scoring logic presented in this section. Figure 3 summarizes the overall pipeline, and Figure 4 presents the pillar–subpillar–indicator hierarchy.

5.1 Indicator normalization

Indicators are coded from verifiable artifacts or external evaluations and mapped to $[0, 1]$. Binary indicators map Yes $\mapsto 1$, No $\mapsto 0$, and *Unknown* to missing. Ordinal indicators use a generic rubric mapping $\{0, 1, 2\} \mapsto \{0, 0.5, 1\}$. For counts c we use a tempered logarithmic mapping

$$s(c) = \min\left(1, \frac{\log(1 + c)}{\log(1 + c_{\text{ref}})}\right), \tag{1}$$

where c_{ref} is the 95^{th} percentile within the release. The transform is monotone, bounded, less sensitive to outliers than linear scaling, and comparable across releases because c_{ref} is frozen per version.

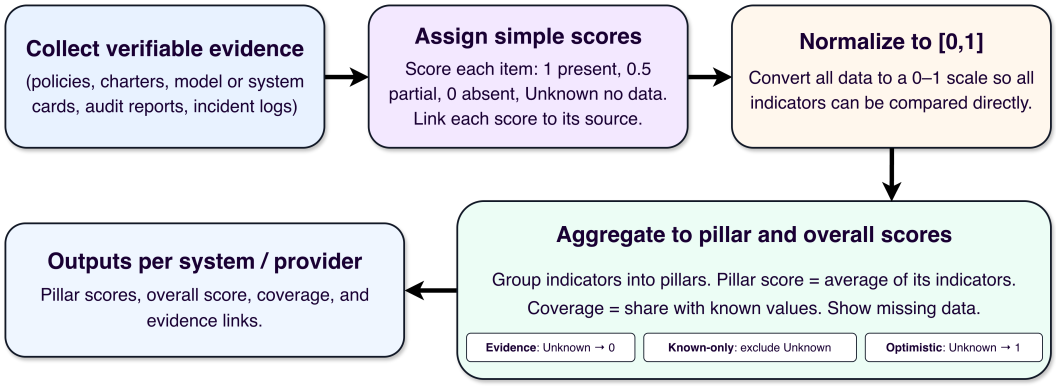

Figure 2: Simple schematic of the AIPI measurement model. Indicators are coded from verifiable artifacts, normalized, aggregated to pillar means, and reported as an interval via explicit treatments of *Unknown* (evidence, known-only, optimistic) with coverage.

5.2 Aggregation and properties

Let i index systems (or producers) and $p \in \{\mathrm{PG}, \mathrm{ID}, \mathrm{TR}, \mathrm{AC}\}$ index pillars. For indicator set $\mathcal{J}_p$, the *known-only* pillar score averages normalized values s_{ijp} over indicators with evidence:

$$S_{ip}^{\mathrm{known}} = \frac{1}{|\mathcal{J}_{ip}^{\mathrm{known}}|} \sum_{j \in \mathcal{J}_{ip}^{\mathrm{known}}} s_{ijp}. \tag{2}$$

The *evidence* model treats Unknown as 0:

$$S_{ip}^{\mathrm{evid}} = \frac{1}{|\mathcal{J}_p|} \sum_{j \in \mathcal{J}_p} s_{ijp}^{*}, \qquad s_{ijp}^{*} = \begin{cases} s_{ijp} & \text{if known,} \\ 0 & \text{if Unknown.} \end{cases} \tag{3}$$

Coverage is reported as $C_{ip} = |\mathcal{J}_{ip}^{\mathrm{known}}|/|\mathcal{J}_p|$. Overall scores average the pillars:

$$\mathrm{AIPI}_i^{\bullet} = \frac{1}{4} \sum_{p} S_{ip}^{\bullet}, \qquad \bullet \in \{\mathrm{known}, \mathrm{evid}\}. \tag{4}$$

A provider's score averages across its K systems:

$$\mathrm{AIPI}_{\mathrm{provider}}^{\bullet} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{AIPI}_{i_k}^{\bullet}. \tag{5}$$

Bounding and reporting. An upper bound treats Unknown as 1:

$$S_{ip}^{\mathrm{opt}} = \frac{1}{|\mathcal{J}_p|} \sum_{j \in \mathcal{J}_p} s_{ijp}^{+}, \qquad s_{ijp}^{+} = \begin{cases} s_{ijp} & \text{if known,} \\ 1 & \text{if Unknown.} \end{cases} \tag{6}$$

For every system and pillar: $S_{ip}^{\mathrm{evid}} \le S_{ip}^{\mathrm{known}} \le S_{ip}^{\mathrm{opt}}$. The reporting convention is to publish $(\mathrm{AIPI}_{\mathrm{evid}}, \mathrm{AIPI}_{\mathrm{known}}, \mathrm{AIPI}_{\mathrm{opt}})$ as a closed interval with a marked midpoint, plus pillar-level coverage intervals $\left(C_{ip}^{\min}, C_{ip}^{\max}\right)$ where the minimum treats un-attributable evidence as uncovered and the maximum includes neutral evidence verified by third parties (Sec. 15).

Estimator guidance. For procurement and safety-critical use, we recommend $\mathrm{AIPI}_{\mathrm{evid}}$ (the lower-bound score) with minimum floors for the overall score, each pillar S_{ip}^{evid}, and coverage C_{ip} (by pillar and on average). As a basic check, confirm that the provider publishes: a vulnerability disclosure policy and handling process, a redress channel with stated response times, and a current model/system card describing purpose, data governance, intended uses/limits, and incident history. For research comparisons and longitudinal tracking, report the known-only midpoint $\mathrm{AIPI}_{\mathrm{known}}$ together with the full interval $[\mathrm{AIPI}_{\mathrm{evid}}, \mathrm{AIPI}_{\mathrm{opt}}]$ and pillar-level coverage.

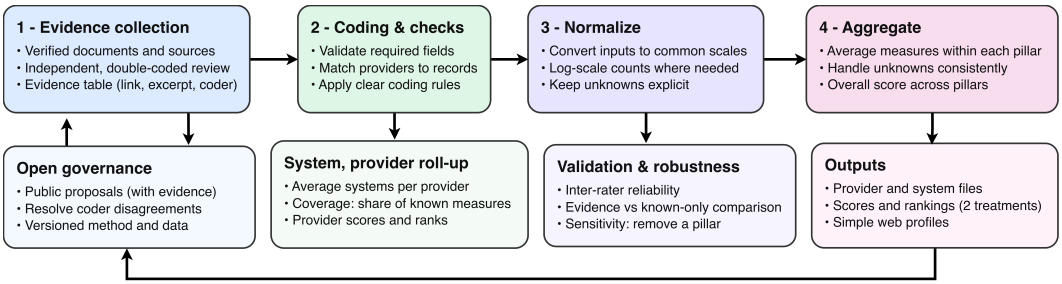

Figure 3: AIPI methodology. Verifiable artifacts, structured web & repository analysis, external evaluations, and expert interviews feed a coded dataset with evidence links. After validation and normalization, pillar scores are aggregated under two treatments of Unknown (*evidence* vs. *known-only*) to produce system- and provider-level AIPI. Validation reports inter-rater reliability, cross-index correlations, and sensitivity.

5.3 Uncertainty and reliability

Uncertainty arises from coding disagreement and incomplete evidence. Each release double-codes a stratified sample and reports inter-rater reliability and documents sampling decisions [10, 24]. Sensitivity analysis varies indicator inclusion and performs leave-one-pillar-out checks; rank stability is summarized using Kendall's τ between alternative score vectors [10]. Release notes publish coverage distributions and rank intervals.

6 Indicators and coding guide

AIPI uses a public codebook that groups indicators into four pillars: participatory governance, inclusivity and diversity, transparency, and accountability. Each indicator must be supported by a public artifact or vetted external evaluation and is scored on a common $[0, 1]$ scale. Release notes document edge cases and adjudication precedents.

Operationalization. Participatory-governance indicators capture documented stakeholder input linked to dispositions; inclusivity indicators capture representation, language access, accessibility, and subgroup evaluation; transparency indicators capture cards, datasheets, release notes, and governance documentation; accountability indicators capture incident handling, disclosure processes, audits, redress, and remediation. Full indicator text and coding examples are provided in the public codebook and appendix.

7 Cohort, eligibility, discovery, and timelines

Eligibility. We included a producer if, by the cutoff date, it met all of the following: (i) develops or maintains a general-purpose foundation model or a generative AI service that is widely available; (ii) controls releases, governance, or deployment guardrails; and (iii) has at least one system family or API that is publicly identifiable. We applied the same criteria to government consortia and non-profit organizations.

Provider discovery. We identified candidate producers using four sources: (1) Google search and news coverage; (2) academic and industry reports; (3) open-source platforms and model registries; and (4) references in standards documents, audits, and government procurement records.[1]

Timeline and data collection. We set the eligibility cutoff at 30 September 2025. We compiled the dataset over four months, from early May 2025 to late September 2025. During this period, we collected and timestamped evidence for each producer.

[1]We did not collect any personal data.

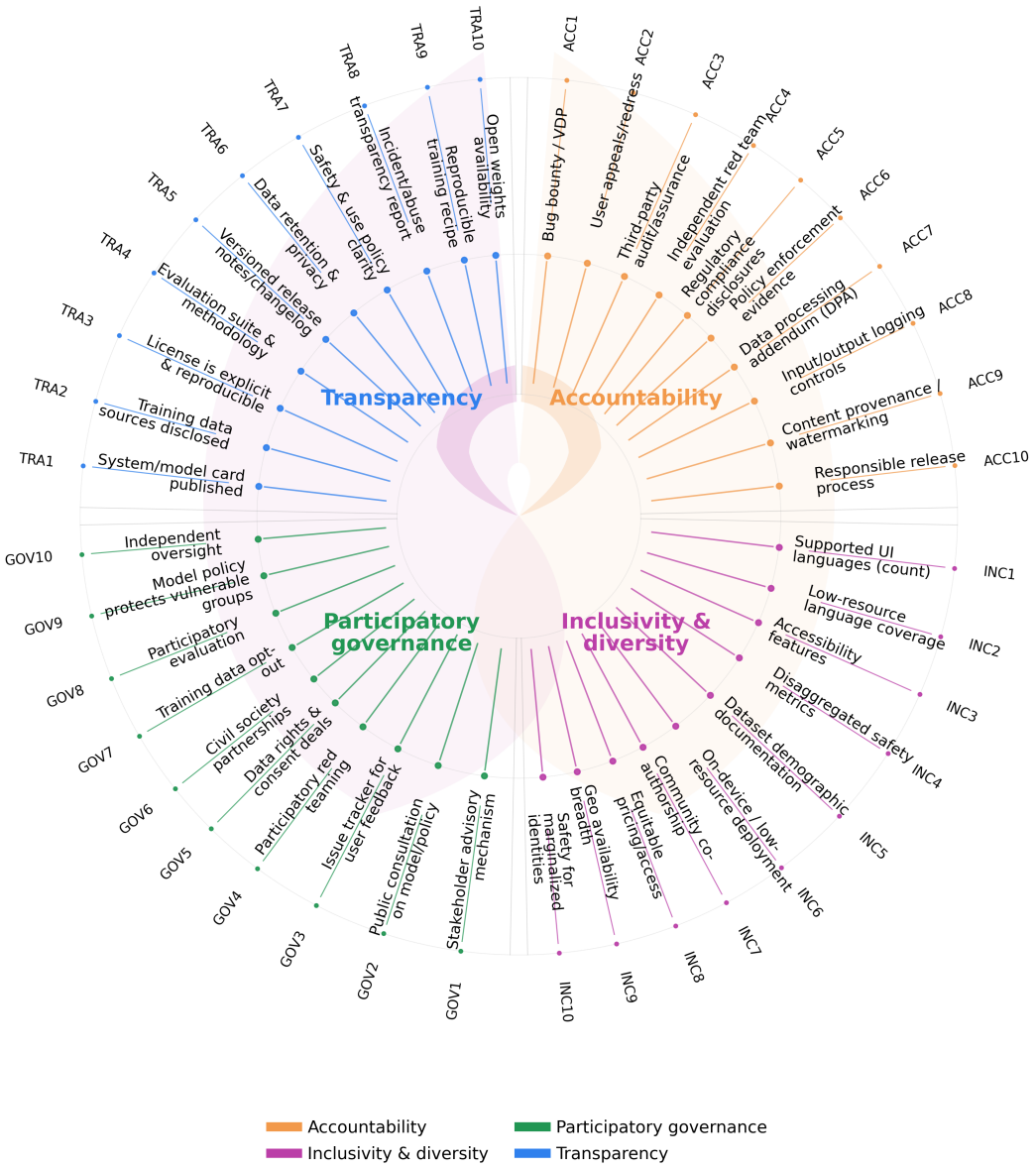

Figure 4: Pillar–subpillar–indicator hierarchy. The radial diagram enumerates indicators within each pillar and subpillar; the outer band lists indicator IDs.

8 Data sources and pipeline

The pipeline integrates three streams. First, structured web and repository search traverses producer sites, documentation hubs, transparency reports, model and dataset cards, and registries; enumeration uses Google search and curated news articles to seed candidate providers, then applies de-duplication and eligibility filters (Sec. 7). Reproducibility metadata include query templates, URLs, timestamps, and retrieval logs. Second, external independent evaluations and benchmarks are incorporated with explicit citation, for example large-scale capability benchmark syntheses and public safety assessments [35, 2]. Third, expert interviews with researchers, journalists, and civil society practitioners provide context for interpreting artifacts and identifying missing documentation; interviews obtain informed consent and exclude sensitive data. Evidence links are checked for accessibility and archived when permitted. The evidence graph ties producers to indicators, sources, and coders; adjudication logs are public. All builds are deterministic and versioned.

9 Validation

Validation proceeds on three axes. Content validity is established by triangulating the taxonomy against recognized frameworks that articulate participation, inclusion, transparency, and accountability [12, 40, 17, 23]. Construct validity is probed by testing expected correlations: transparency pillar scores should correlate with external transparency syntheses without complete overlap, inclusivity indicators on language and accessibility should correlate with independent coverage measures, and accountability indicators should align with the presence of disclosure programs and audits. Reliability is addressed through coder training, a detailed codebook with exemplars and counter-examples, double coding of a stratified sample each release, and public adjudication. We report inter-rater agreement and document sampling, and target acceptable thresholds for publication [10]. Sensitivity analysis varies pillar inclusion and indicator weights and reports rank stability (e.g., Kendall's τ). To bound uncertainty from missing evidence we publish the evidence and known-only scores plus coverage; for procurement use we recommend minimum thresholds using the lower-bound evidence score.

10 Results

Figure 5 summarizes provider scores under the known-only treatment for an initial cohort. Bars show each pillar's contribution to the overall AIPI (each pillar divided by four). The right column reports coverage, the share of indicators with evidence-coded values.

In this pilot cohort, OpenAI scores 0.32 with the highest coverage ($\approx 40\%$). Anthropic follows at 0.20 (coverage $\approx 28\%$). Stability AI, Mistral AI, and Google DeepMind cluster between 0.15 and 0.17, while BigScience, EleutherAI, UNICEF Office of Innovation, Climate TRACE, Meta, and Alibaba Cloud are near 0.14; VinAI registers 0.12. Coverage ranges from roughly 15% to 40% with a median near 18%. Within-provider variation across pillars is informative: some producers obtain most credit from transparency while lagging in participatory governance or accountability, suggesting opportunities for targeted improvement. Because known-only scores exclude missing indicators, we report the evidence scores alongside coverage in the release dataset; the difference between the two reflects documentation gaps rather than practice alone.

11 Governance, implementation, and releases

AIPI is implemented as a public repository with deterministic builds and archived artifacts. Anyone can file an issue to challenge a code, propose an indicator change, or contribute evidence. Issues are triaged by volunteer reviewers; disagreements are logged and escalated to a multi-stakeholder maintainer council spanning research, civil society, industry, and regions. Each release is versioned and archived with a change log. The website provides system and provider profiles, an API, and download links for evidence tables and scores. Governance adheres to the same pluralistic principles that the instrument evaluates and aligns with open practices in accessibility and security communities [17, 25].

We operationalize this governance with three layers of decision making. First, a release team executes evidence discovery, coding, and quality assurance using a frozen codebook for that version. Second, an adjudication panel resolves disputed codes using a publicly documented standard of evidence, logging rationales and precedents so that similar cases are treated consistently across releases. Third, the maintainer council decides on indicator additions, deprecations, and major rubric changes through a transparent proposal process that includes (i) an open call for change requests, (ii) a public comment window, and (iii) recorded decisions with an explicit migration plan for longitudinal comparability.

To address concentration and geographic skew in participation, the maintainer council is constituted to include representation across stakeholder categories and regions (research, civil society, industry, labor, and policy), with time-bounded terms and conflict-of-interest disclosures. Members recuse themselves from decisions involving their own organizations. Releases follow a predictable cadence with an evidence freeze date, double-coding of a stratified sample, and a short pre-release audit period during which external contributors can contest evidence links or propose missing artifacts.

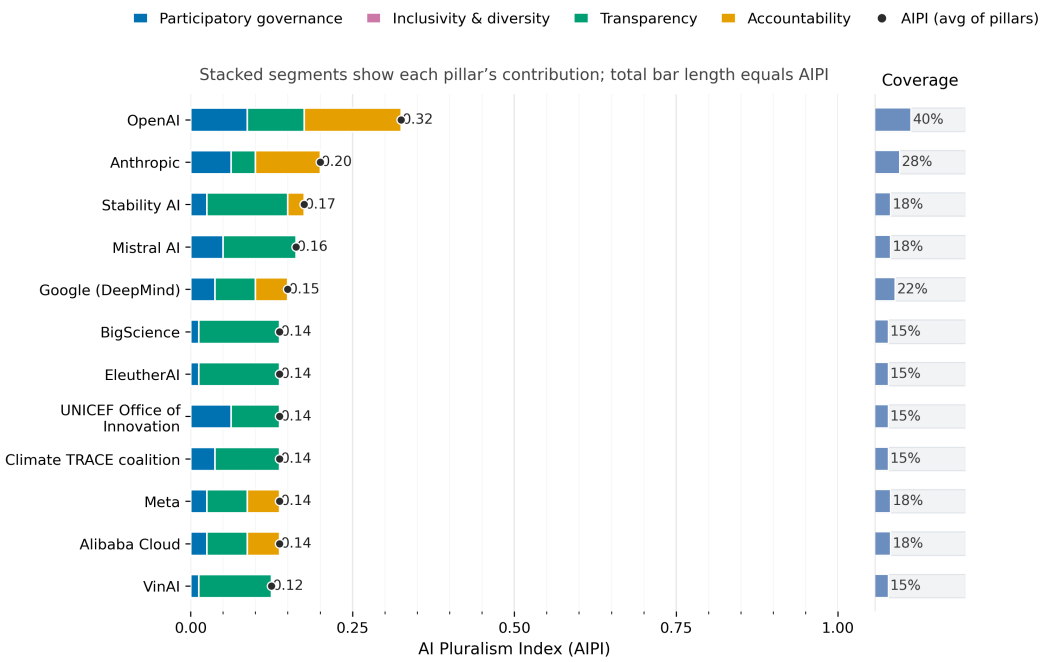

Figure 5: Provider results, known-only treatment. Stacked segments show each pillar's contribution (each pillar divided by four); total bar length equals the AIPI value on a 0–1 scale. The right column reports coverage, the share of indicators with evidence-coded values.

Final releases are archived, and we publish a changelog that distinguishes schema changes (which affect comparability) from evidence updates (which affect scores).

12 Recommendations for practice and policy

For producers, a practical path to improving AIPI is to publish verifiable artifacts that reflect real processes, and to ensure that those processes include affected stakeholders across deployment contexts. Model and system cards should disclose data governance, known limitations, intended uses, subgroup performance, and update histories. Vulnerability disclosure processes should accept reports from good-faith researchers and specify service levels; vulnerability handling should name a PSIRT (Product Security Incident Response Team) function and lay out triage and remediation steps. Public consultations should summarize input and publish a disposition matrix that links proposals to decisions and downstream changes, rather than collecting comments without response. Where systems are deployed internationally, participatory mechanisms should be accessible across regions, for example via remote participation, multilingual materials, and geographically diverse advisory membership.

For procurers, AIPI can be used as a practical checklist rather than a headline ranking. Reference the indicator list directly in requests for proposals; require minimum evidence scores and explicit coverage thresholds; and ask vendors to supply the primary artifacts (not summaries) referenced by their scores. Use pillar-level scores to tailor due-diligence: for low participatory-governance scores, require a roadmap for consultation and redress; for low transparency scores, require current system cards and change logs; for low accountability scores, require incident response and audit documentation.

For regulators and standards bodies, several indicators map onto emerging governance requirements, including documentation, risk management, incident reporting, accessibility, and redress. AIPI can support supervision by translating these requirements into auditable evidence checks and by highlighting where documentation is absent, stale, or geographically narrow. Regulators can also use AIPI's coverage reporting to prioritize oversight, focusing audits where low scores are paired with low evidence coverage.

For researchers and civil society, AIPI's evidence graph enables systematic scrutiny of claims and targeted advocacy that rewards concrete improvements. We encourage third-party replications and "shadow scores" for contested providers, and we treat well-documented replications as admissible evidence in subsequent releases.

13 Limitations

AIPI relies on observable artifacts and external evaluations, so it may undercount non-public internal practices and could be gamed if documentation is published without corresponding practice. We mitigate this by requiring time-stamped artifacts, crediting external audits and incident disclosures, and publishing coverage to expose documentation gaps. The index inherits biases from English-language documentation; we prioritize multilingual evidence and weight accessibility and language coverage. We do not measure capability or risk directly and should be used alongside capability, safety, and domain-specific evaluations [35, 6, 2]. Some sources are journalistic or blog posts; these are treated as context, with core claims anchored to primary or official texts where possible.

A further limitation concerns scope and survivorship bias. Because the unit of analysis is the producer and ongoing system families, many participatory, community-led "plural-AI" initiatives (often experimental or inactive) are underrepresented. This risks overweighting well-resourced providers and understating pluralistic innovation. Future releases will pilot a community-maintained registry and evidence track for plural-AI projects, including non-service and archived initiatives meeting minimal provenance requirements. Crowdsourced nominations and review, integrated into adjudication and versioning, would expand coverage and improve discoverability. Results from this track will be reported separately to avoid conflating prototypes with production systems.

Finally, the cohort is expanding; early comparisons should be interpreted as a baseline rather than a definitive ranking.

14 Indicators and coding guide

Indicators are designed to be *observable, verifiable, and consequential*. Each code must be supported by a link to a public artifact (policy, governance record, model/system card, audit report, standard conformance statement) or a vetted external evaluation. Detailed rubrics and edge cases are available in the public codebook on GitHub: `https://github.com/rsdmu/aipi-pluralism-index`. For additional details on the methodology, indicator construction, and validation process, please consult the arXiv version of this paper [28].

15 Future work

Next releases will (i) add multilingual evidence discovery protocols, including targeted search in major languages and structured prompts for local documentation; (ii) reward third-party local verifications, privileging evidence from independent community groups and regional agencies; and (iii) document regional coverage explicitly by tagging artifacts with geography and publishing pillar-level coverage intervals by region. We will release partial-dependence plots of AIPI versus coverage and report where rank flips occur; policy guidance will recommend $\mathrm{AIPI}_{\mathrm{evid}}$ for procurement and $\mathrm{AIPI}_{\mathrm{known}}$ with intervals for comparative research. We will also extend correlations against transparency syntheses, capability facets, and safety reports to the growing cohort and publish bootstrapped confidence intervals.

16 Conclusion

Pluralistic AI requires structures that allow many voices to shape technical and governance choices, as well as mechanisms that make those structures visible and accountable. The AI Pluralism Index provides a principled, auditable, and adaptable way to measure progress toward that goal. We will publish regular releases that include evidence graphs, adjudication logs, reliability statistics, and sensitivity analyses. These resources provide open access to the framework, documentation, and implementation, enabling replication, scrutiny, and continued community contributions.

References

[1] Sherry R. Arnstein. A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4):216–224, 1969. doi: https://doi.org/10.1080/01944366908977225.

[2] Yoshua Bengio et al. International scientific report on the safety of advanced ai (interim report), 2024. URL https://arxiv.org/abs/2412.05282.

[3] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006. URL http://www.jstor.org/stable/j.ctt1njknw.

[4] James Bohman. *Public Deliberation: Pluralism, Complexity, and Democracy*. MIT Press, 2000.

[5] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023. URL https://arxiv.org/abs/2310.12941.

[6] Rishi Bommasani et al. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

[7] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024. URL https://arxiv.org/abs/2403.04132.

[8] Council of Europe. The framework convention on artificial intelligence and human rights, democracy and the rule of law. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence, 2024.

[9] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

[10] J. W. Creswell. Qualitative inquiry and research design: Choosing among five approaches. SAGE., 2013.

[11] European Union. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 2024.

[12] European Union. Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 2024.

[13] Batya Friedman and David G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019. URL https://mitpress.mit.edu/9780262039536/value-sensitive-design/.

[14] Archon Fung. Varieties of participation in complex governance. *Public Administration Review*, 66(s1):66–75, 2006. doi: 10.1111/j.1540-6210.2006.00667.x.

[15] Future of Life Institute. 2025 ai safety index (summer 2025). https://futureoflife.org/ai-safety-index-summer-2025/, 2025.

[16] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64 (12):86–92, November 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

[17] G. Goggin, K. Ellis, and W. & Hawkins. Disability at the centre of digital inclusion: Assessing a new moment in technology and rights. Communication Research and Practice, 5(3), 290–303. https://doi.org/10.1080/22041451.2019.1641061, 2019.

[18] Thilo Hagendorff. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, 2020. doi: 10.1007/s11023-020-09517-8. URL https://link.springer.com/article/10.1007/s11023-020-09517-8.

[19] International Organization for Standardization. Iso/iec 29147:2018 information technology—security techniques—vulnerability disclosure. https://www.iso.org/standard/72311.html, 2018.

[20] International Organization for Standardization. Iso/iec 30111:2019 information technology—security techniques—vulnerability handling processes. https://www.iso.org/standard/69725.html, 2019.

[21] International Organization for Standardization. Iso/iec 42001:2023 artificial intelligence management system (aims). https://www.iso.org/standard/42001, 2023.

[22] Prithvi Iyer. The foundation model transparency index: What changed in 6 months? https://techpolicy.press/the-foundation-model-transparency-index-what-changed-in-6-months/, May 2024.

[23] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1:389–399, September 2019. doi: 10.1038/s42256-019-0088-2. URL https://doi.org/10.1038/s42256-019-0088-2.

[24] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4 edition, 2018.

[25] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4):611–627, December 2018. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-017-0279-x. URL http://link.springer.com/10.1007/s13347-017-0279-x.

[26] D. Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector. Zenodo. https://doi.org/10.5281/ZENODO.3240529, 2019.

[27] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596, 2019.

[28] Rashid Mushkani. Measuring what matters: The ai pluralism index, 2026. URL https://arxiv.org/abs/2510.08193.

[29] Rashid Mushkani, Hugo Berard, Toumadher Ammar, Cassandre Chatonnier, and Shin Koseki. Co-producing ai: Toward an augmented, participatory lifecycle. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1785–1799. AAAI/ACM, 2025. doi: 10.1609/aies.v8i2.36674.

[30] Rashid Mushkani, Hugo Berard, Allison Cohen, and Shin Koseki. Position: The right to ai. *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://arxiv.org/abs/2501.17899.

[31] Wilhelmina Nekoto et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.findings-emnlp.195.

[32] OECD. Recommendation of the council on artificial intelligence. *OECD Legal Instruments*, 2019. URL https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

[33] OECD, European Union, and European Commission Joint Research Centre. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris, 2008. doi: 10.1787/9789264043466-en. URL https://doi.org/10.1787/9789264043466-en.

[34] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 33–44, 2020. doi: 10.1145/3351095.3372873.

[35] Partha Pratim Ray. Benchmarking, ethical alignment, and evaluation framework for conversational ai: Advancing responsible development of chatgpt. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(3):100136, 2023. ISSN 2772-4859. doi: https://doi.org/10.1016/j.tbench.2023.100136. URL https://www.sciencedirect.com/science/article/pii/S2772485923000534.

[36] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments report: A practical framework for public agency accountability. AI Now Institute, April 2018. URL https://ainowinstitute.org/reports.html.

[37] Responsible AI Collaborative. Ai incident database. https://incidentdatabase.ai/, 2024.

[38] Stanford Center for Research on Foundation Models (CRFM). Helm capabilities. https://crfm.stanford.edu/2025/03/20/helm-capabilities.html, March 2025.

[39] Stanford CRFM. Holistic evaluation of language models (helm). https://crfm.stanford.edu/helm/, 2024.

[40] Elham Tabassi. AI risk management framework: AI RMF (1.0). Technical report, National Institute of Standards and Technology (NIST), 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[41] S. Umbrello and I. van de Poel. Mapping value sensitive design onto ai for social good principles. AI and Ethics, 1(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3, 2021.

[42] UNESCO. Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455, 2021.

[43] World Wide Web Consortium (W3C). Web content accessibility guidelines (wcag) 2.2. https://www.w3.org/TR/WCAG22/, 2023.