

Online Resource 1: Blinded Technical Appendix

Evidence for Whom? Contestability and Socio-Technical Assurance in High-Risk AI Governance

Synchronized to the deposited data-and-code bundle; author identities withheld

1. Purpose and status

This appendix sets out the crosswalk, the reliability analysis, the sensitivity checks, and the legal-scope boundaries that the article summarizes. It is a companion, not a second manuscript, so it does not repeat the article's theoretical argument. The machine-readable artifacts and the scripts that produce them are in Online Resource 2. Throughout, a crosswalk code records technical relevance and says nothing about legal conformity.

2. Evidence set and search limits

The evidence set is purposive and citation-verified. It is not an exhaustive or reproducible scoping-review corpus: the original per-database yields and the full candidate-to-exclusion ledger were not retained. Figure 1 in the article is therefore a historical accounting assembled from project records, not a PRISMA flow diagram, and no PRISMA diagram can be produced for a search that cannot be reconstructed.

Pool	Memberships	Qualification
Indexed records	147	Includes two official NIST documents; not all rows are technical methods
Reused fairness supplement	202	Used only to deepen M2; record counts do not set cell codes
Official texts	5	NIST 100-1; NIST 600-1; EU 2024/1689; ISO/IEC 42001; ISO/IEC 23894
Total memberships	354	Pools overlap
Unique sources	347	Verified after canonical deduplication

The canonical manifest identifies the seven duplicate memberships. NIST AI 100-1 accounts for three: it is indexed twice, once under M1 and once as a governance record, and also appears in the fairness and official pools. NIST AI 600-1, RealToxicityPrompts, Beyond Memorization, and Regulation (EU) 2024/1689 account for one each.

3. Crosswalk unit and codebook

Rows are twelve method families. Columns are 22 selected entries: nine NIST functions or subcategories, seven EU AI Act articles, and six ISO/IEC 42001 control groupings. These columns are not normalized units, and a broad or omnibus provision may attract more families simply because it contains more evidentiary facets. Breadth does not, however, drive the result: columns with no primary family engage 10.8 method families on average through some code, against 9.0 for columns with two or more primaries, so on this measure the columns without a chief evidence source are the broader ones (tools/build_granularity_check.py).

Code	Label	Meaning
D	Primary	The method output is the evidence the selected requirement chiefly asks for
P	Supporting	The output contributes but is insufficient alone
I	Contextual	The output informs interpretation without serving directly as evidence
blank	Not applicable	No material technical relevance under this coding

3.1 NIST AI RMF panel

Family	GV	MAP	M2.3	M2.6	M2.7	M2.9	M2.10	M2.11	MG
M1	I	P	D	I				I	P
M2	I	P	I	P				D	P
M3	I	P	P	P	D				P
M4	I	P	I	D	P		I	I	P
M5	I	P		D	D		I	I	P
M6	I	I		I		D		P	I
M7		I	P	P		I			P

Family	GV	MAP	M2.3	M2.6	M2.7	M2.9	M2.10	M2.11	MG
M8	I	D	I				I	P	I
M9	I	P	I			I	I	I	I
M10	I	I		P		P		I	P
M11	I	I	I	P	P		I	I	D
M12	I	P		I	P		D	I	P

3.2 EU AI Act panel

Family	A9	A10	A11	A12	A13	A14	A15
M1	P	I	P		P		D
M2	P	P	P		P	I	P
M3	P	I	P		I		D
M4	P		P		I	P	P
M5	P		P			I	D
M6	I		P		P	P	
M7	P		P		P	P	P
M8	I	D	P		P		P
M9	I	I	D	I	D	P	
M10	I		I		P	D	
M11	P		I	D	I	I	P
M12	P	P	P		P		P

3.3 ISO/IEC 42001 panel

Family	IA5	IA6	IA7	IA8	IA626	IA10
M1	I	D	I	P	I	I
M2	P	P	P	P	P	I
M3	I	D	I	I	P	
M4	P	D		P	P	I
M5	P	D		I	P	I
M6	I	P		P		
M7	I	P		P	P	
M8	P	I	D	P		P
M9	P	I	P	D	I	P
M10	P	P		P	P	
M11	P	P	I	I	D	P
M12	P	P	P	I	I	I

4. Reliability and adjudication

The frozen matrix immediately before either 80-cell non-author rater was incorporated is released as mapping_matrix_pre_independent_validation.csv. This permits uncontaminated comparison. Both raters used standardized author-prepared method and requirement descriptions and did not see the authors' codes nor each other's. The exercise does not validate search completeness, family assignment, legal interpretation, the socio-technical assurance framework, or the gap definition.

Comparison	n	Exact	Unweighted κ	Weighted κ	AC2	Primary/not	Binary κ
Frozen matrix vs technical rater	80	57.5%	0.43	0.73	0.79	88.8%	0.68
Frozen matrix vs legal rater	80	53.8%	0.38	0.67	0.73	88.8%	0.70
Technical vs legal rater	80	72.5%	0.58	0.82	0.92	n/a	n/a
Post-rule matrix vs technical rater	80	66.2%	0.55	0.77	0.82	97.5%	0.93
Post-rule matrix vs legal rater	80	62.5%	0.50	0.71	0.77	97.5%	0.93

The last two rows are post-adjudication descriptions, not independent confirmation. The consensus rule was post hoc but applied symmetrically: where both non-author raters crossed the primary/not-primary boundary in the same direction, the post-rule matrix adopted their shared judgment. It removed four primaries and added three. The named authors applied that rule themselves. Three language-model adjudicators then re-read the seven cells from a blinded packet, confirming six and reversing E233, which restored M11 as primary for Article 12. All three matrices are therefore deposited, the model-adjudicated one as the governing specification, and the article's qualitative conclusions hold under each. Bootstrap 95% intervals for the frozen-reference weighted coefficients are [0.61, 0.82] and [0.53, 0.77]. Unweighted and quadratically weighted coefficients are reported together because they differ substantially: on a four-level ordinal scale where 95% and 91% of disagreements are adjacent, quadratic weighting discounts most of the observed disagreement, so the weighted figures should not be read as the agreement rate.

4.1 Recoding-sample family coverage

Family	Sampled cells	Family	Sampled cells	Family	Sampled cells
M1	10	M5	9	M9	0
M2	7	M6	9	M10	8
M3	9	M7	5	M11	7
M4	7	M8	2	M12	7

M9 was absent and M8 had only two cells. That shortfall is structural rather than incidental: within each stratum the sampler ordered cells by family label and then selected at a fixed stride, truncating at twenty, so the families sorting last, M8 and M9, fell in the discarded tail. The five frozen primary cells no rater saw are exactly that tail: M8 x EU Art. 10, M8 x ISO A.7, M9 x EU Art. 11, M9 x EU Art. 13, and M9 x ISO A.8. Those five cells are not unexamined, though. The within-team cross-check coded all 264 cells blind to the author matrix, M8 and M9 included, and it codes all five of them primary; across the 44 M8/M9 cells it agrees with the post-rule matrix on 97.7% of the primary-versus-not boundary (tools/build_m8m9_check.py). Being within-team, that check is weaker evidence than the non-author exercise. Reliability from the 80-cell sample remains an aggregate stratified estimate and does not justify treating any single primary cell as individually reproduced. The sampler and the blank rater instrument are in Online Resource 2, so this can be checked directly.

5. Post-rule coverage reading and sensitivity

Framework	Requirement	Primary families	Coverage
NIST	GV	0	Narrow
NIST	MAP	1	Moderate
NIST	M2.3	1	Moderate
NIST	M2.6	2	Moderate
NIST	M2.7	2	Moderate
NIST	M2.9	1	Moderate
NIST	M2.10	1	Moderate
NIST	M2.11	1	Moderate
NIST	MG	1	Moderate
EU	A9	0	Narrow
EU	A10	1	Moderate
EU	A11	1	Moderate
EU	A12	1	Moderate
EU	A13	1	Moderate
EU	A14	1	Moderate
EU	A15	3	Broad
ISO	IA5	0	Narrow
ISO	IA6	4	Broad
ISO	IA7	1	Moderate
ISO	IA8	1	Moderate
ISO	IA626	1	Moderate
ISO	IA10	0	Narrow

Four columns have no primary family in the governing matrix: NIST GOVERN, EU Article 9, and ISO A.5 and A.10. Fourteen of sixteen moderate columns rest on one primary cell. Under either non-author rater, the remaining zero-primary columns stay narrow; A.5 changes only under the within-team legal cross-check and are therefore reported as sensitivity, not independent recoding. Article 12 is especially coding-sensitive: its only frozen primary, M11, was changed to supporting when both non-author raters crossed the primary/not-primary boundary, and the model re-adjudication then restored it. NIST GOVERN, Article 9, and ISO A.5 and A.10 have no primary family at any stage. Seven of the 22 columns change primary count between the frozen and author post-rule matrices, and one further column, Article 12, changes between that stage and the governing matrix.

6. Legal-scope boundaries

- Article 12 requires automatic event logging and traceability appropriate to intended purpose; it does not expressly require individual-decision replay or affected-person access. Reconstruction adequacy is the authors' normative proposal.
- Article 14 allocates design and implementation aspects across providers and deployers and concerns effective oversight during use; it is not itself an individual right of recourse.
- Article 27 applies only to specified deployers and uses and is outside the 22 coded columns.
- For Annex III points 2–8, Article 43(2) uses Annex VI internal control without routine notified-body examination. Point 1 and the Annex I Section A product routes require separate qualification.
- Articles 74, 77, 79, 85, and 86 preserve post-market, authority, complaint, explanation, and corrective mechanisms; the article does not claim evidence is never independently examined.
- Article 40's presumption of conformity extends only to requirements covered by a harmonized standard whose reference is published in the Official Journal. Formal non-industry participation routes exist under Regulation (EU) 1025/2012.

7. Reproduction from Online Resource 2

After extracting the ZIP into an empty directory, run:

```
python tools/build_source_manifest.py
python tools/build_matrix.py
python tools/make_figures.py
python tools/build_agreement_tables.py
python tools/build_sensitivity_table.py --check
python tools/build_m8m9_check.py
python tools/build_granularity_check.py
```

The package contains the frozen and post-rule matrices, the adjudication ledger, the canonical source manifest, derived agreement tables, generated figures, and dual-license terms. The two non-author raters' cell-level ledgers are included with each rater's written consent; their free-text comment columns are removed, because neither confirmation form confirmed those comments' accuracy.

corpus/agreement_tables.csv gives the 4x4 contingency table for every reported comparison, and tools/build_agreement_tables.py recomputes exact agreement, Cohen's kappa, quadratically weighted kappa and the binary primary-versus-not figures from those tables alone, so every statistic in the article is reproducible from this package. Successful regeneration does not repair the unavailable historical search path.

8. Limitations carried into the main article

- single-reviewer screening and an unreconstructable historical search path;
- non-disjoint evidence pools and unequal depth across method families;
- columns of unequal breadth and institutional function, though breadth alone does not explain the coverage pattern;
- taxonomy and socio-technical lens developed through engagement with the same materials;
- aggregate reliability, with no M9 cells and only two M8 cells in the 80-cell sample, partly offset by the within-team cross-check of all 264 cells;
- no affected-person or advocate participation;
- English-only and Global-North instrument scope;
- coverage codes indicate relevance, not method quality, evidentiary sufficiency, or legal conformity.