

A pre-registered, falsifiable measurement-methodology bridge from styxx primitives to NIST AI RMF 1.0 Measure-function subcategories (v0.1)

Author: Alexander Rodabaugh (Fathom Lab) **Date:** 2026-05-28 **Substrate:** styxx 7.7.10 in-development; companion to `papers/EU_AI_ACT_COMPLIANCE_2026.md` (EU AI Act bridge v0.1) and `papers/PAPER_recursive_discipline_2026_05_27.md` (recursive-discipline methodology paper v7); module `styxx.compliance.nist_ai_rmf` ships in the same commit **Status:** v0.1 minimum-viable mapping. Companion to the EU AI Act bridge, applying the same methodology to a parallel jurisdictional regime (United States NIST AI RMF 1.0). Not legal advice; not a substitute for an organization’s own NIST AI RMF implementation.

Abstract

NIST AI 100-1 (the AI Risk Management Framework 1.0, January 2023) defines four core functions for AI risk management: Govern, Map, Measure, Manage. The Measure function is its analytical engine — 22 subcategories that evaluate AI systems against seven trustworthy characteristics (valid/reliable, safe, secure/resilient, accountable/transparent, explainable/interpretable, privacy-enhanced, fair/bias-mitigated). The RMF is voluntary in US federal contexts but is increasingly referenced by federal procurement, state-level legislation, and private-sector contracting. This paper introduces `styxx.compliance.nist_ai_rmf`, a parallel jurisdictional bridge to the EU AI Act compliance bridge introduced the same day. It maps styxx primitives to five Measure subcategories (MS-2.3, MS-2.4, MS-2.5, MS-2.6, MS-2.13) with calibrated metrics, construct ceilings, commit-level reproducibility receipts, and the same three kill-gates (A1 specific subcategory citations, A2 construct ceilings disclosed, A3 uncovered covered) enforced by tests. v0.1 explicitly enumerates six uncovered Measure subcategories (MS-2.7, MS-2.8, MS-2.9 partial, MS-2.10, MS-2.11, MS-2.12) with alternative-tool references. The bridge shares dataclasses with the EU AI Act bridge through `styxx.compliance._common`, signaling that the methodology pattern transfers across regulatory regimes. Not legal advice; independent review required for production deployment.

1. Background

1.1 NIST AI RMF 1.0 in context

NIST AI 100-1, published January 2023, organizes AI risk management into four functions: **Govern** (organizational policies), **Map** (context establishment, risk identification), **Measure** (analytical evaluation), **Manage** (risk prioritization and response). The Measure function operationalizes the framework’s analytical claims via 22 subcategories distributed across four categories (MS-1.x, MS-2.x, MS-3.x, MS-4.x). MS-2.x specifically addresses the seven trustworthy AI characteristics enumerated in NIST AI 100-1 §4: valid and reliable; safe; secure and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; fair with harmful bias managed.

The RMF is **voluntary** in US federal contexts. Its adoption is nevertheless meaningful: federal procurement contracts increasingly reference it, state legislatures (notably California, New York, Colorado) cite it in AI governance bills, and major US enterprises adopt it as a de-facto standard pending future statutory frameworks.

1.2 Why a parallel bridge

The EU AI Act bridge (`papers/EU_AI_ACT_COMPLIANCE_2026.md`) addressed one jurisdictional regime (Europe, mandatory, August 2026 enforcement deadline). The NIST bridge addresses a parallel regime (US, voluntary, no fixed deadline) with a structurally different framework: NIST organizes by *function* (Govern/Map/Measure/Manage) and *trustworthy characteristics* (seven), where the EU AI Act organizes by

article and *Annex III high-risk category*. Mapping styxx primitives to *both* frameworks demonstrates that the underlying methodology — pre-registration, construct ceilings, commit-level receipts, kill-gates A1/A2/A3 — generalizes across regulatory regimes.

The bridge is operationally useful: US organizations adopting NIST AI RMF for federal procurement or state-level AI governance can cite the same styxx primitives for Measure subcategories that EU operators cite for Article 15 sub-paragraphs. Cross-jurisdictional AI deployments benefit doubly.

1.3 What this paper IS, and is NOT

IS: a measurement methodology bridge from styxx primitives to NIST AI RMF Measure subcategories, with the same kill-gate discipline as the EU AI Act bridge. v0.1 ships under MIT alongside the underlying primitives. Pre-registered, falsifiable, open-source, commit-level reproducible.

IS NOT: legal advice. A complete NIST AI RMF implementation (the RMF requires Govern, Map, Manage processes that styxx primitives do not address). An endorsement by NIST or any standards body. Sufficient on its own for any contractual or regulatory NIST RMF compliance claim — operators must conduct independent legal and technical review.

2. Methodology

The NIST bridge inherits the methodology from the EU AI Act bridge (see `papers/EU_AI_ACT_COMPLIANCE_2026.md` §2 for the full design principles). Three structural differences:

1. **Citation format:** clauses are MS-N.N (e.g., MS-2.5), not *Article* N.N(x).
2. **Coverage scope:** only the Measure function’s MS-2.x subcategories. The Govern, Map, Manage functions are out of scope by design (styxx is a measurement layer, not a governance/management framework).
3. **Shared dataclasses:** `ComplianceMap`, `PrimitiveCoverage`, `UncoveredRequirement` are imported from `styxx.compliance._common` and shared with the EU AI Act bridge. Both bridges produce the same dataclass instances; only the clause-citation format differs.

The same three kill-gates apply: - **A1 (validity)**. Every clause cites a specific Measure subcategory; regex `^MS-\d+\.\d+$`. - **A2 (falsifiability)**. Every `PrimitiveCoverage.construct_ceiling` is non-empty and at least 50 characters. - **A3 (boundary explicitness)**. `len(uncovered_requirements()) >= len(coverage_table())`. v0.1 ships 6 uncovered vs 5 covered.

A1, A2, A3 are enforced at test time by `tests/test_compliance_nist_ai_rmf.py`.

3. The coverage table

Measure subcategory	Trustworthy characteristic	styxx primitives
MS-2.3 (performance demonstrated qualitatively/quantitatively, measures documented)	valid/reliable	<code>cognometric_card</code> , <code>critique_detector</code> , <code>gauntlet</code> + <code>preflight</code>
MS-2.4 (production monitoring of system functionality)	valid/reliable	<code>cognometric_card</code> , <code>agent_audit</code>
MS-2.5 (valid and reliable; limitations of generalizability documented)	valid/reliable	<code>cognometric_card</code> , <code>critique_detector</code> , <code>gauntlet</code> + <code>preflight</code>

Measure subcategory	Trustworthy characteristic	styxx primitives
MS-2.6 (safety: real-time monitoring, fail-safe under knowledge-limit excursion)	safe	<code>cognometric_card</code> , <code>recover_posture</code> , <code>agent_audit</code>
MS-2.13 (effectiveness of TEVV metrics and processes evaluated)	TEVV process effectiveness	<code>agent_audit</code>

3.1 Why MS-2.5 is the most direct fit

MS-2.5 reads: “*The AI system to be deployed is demonstrated to be valid and reliable. **Limitations of the generalizability beyond the conditions under which the technology was developed are documented.***” (Emphasis added.) The bolded clause is the exact regulatory hook for styxx’s construct-ceiling discipline. Every `PrimitiveCoverage.construct_ceiling` field in the registry documents a published failure mode (e.g., logprob-validity works for refusal but fails for hallucination — confident confabulation; sycophancy FPR 0.30 on restrained-tech responses). These are not buried caveats; they are first-class structured fields enforced by kill-gate A2.

A NIST RMF Measure-function implementer documenting MS-2.5 can cite the styxx primitives + their construct ceilings as direct evidence of generalizability-limit documentation.

3.2 Why MS-2.13 is the recursive subcategory

MS-2.13: “*Effectiveness of the employed TEVV metrics and processes in the measure function are evaluated and documented.*” This is the meta-subcategory — it asks whether the measurement methodology itself is effective.

The styxx project’s `papers/PAPER_recursive_discipline_2026_05_27.md v7` documents six layers of self-falsification of the project’s own measurement methodology, with pre-registered kill-gates and public commit-level receipts. `styxx.agent_audit` is the specific instrument that operationalizes this for substrate-checkable claims (Layer 5 of the arc). The MS-2.13 mapping is unusually clean: the recursive-discipline paper IS the documentation MS-2.13 asks for.

4. What this v0.1 does NOT cover

Per kill-gate A3, this section is intentionally at least as long as §3. Six Measure subcategories are explicitly uncovered:

4.1 MS-2.7 (secure and resilient)

styxx instruments observe agent cognition, not system security or resilience under adversarial conditions. Prompt-injection robustness is *adjacent* (refusal-related instruments may flag some attempts) but security is not the primary scope.

Alternative: Dedicated LLM security tools (Lakera Guard, Prompt-Armor, OWASP LLM Top 10 mitigations); standard application-security audit; penetration testing; red-teaming.

4.2 MS-2.8 (accountable and transparent)

styxx produces measurement evidence; it does not generate deployer-facing accountability documentation, decision-rationale interfaces, or governance artifacts that MS-2.8 implies.

Alternative: Model card frameworks; capability-statement templates; deployer documentation processes; AI ethics review boards.

4.3 MS-2.9 (explainable and interpretable) — partial gap

styxx’s construct-ceiling discipline contributes to interpretability by documenting WHEN each primitive fails, but does not generate per-decision feature attributions, counterfactual explanations, or per-input rationales that modern XAI methods provide.

Alternative: Dedicated XAI libraries (SHAP, LIME, Captum, Inseq); mechanistic interpretability research methods; user studies.

4.4 MS-2.10 (privacy-enhanced)

styxx does not measure differential privacy guarantees, membership inference resistance, training-data memorization, or PII leakage detection. Cognometric instruments operate on outputs but not on privacy properties of those outputs.

Alternative: Differential privacy auditing tools (Google’s DP libraries); PII leakage scanners (Presidio, scrubadub); membership inference test suites.

4.5 MS-2.11 (fair with harmful bias managed)

styxx primitives operate on per-step agent cognition signals; they do not measure population-level disparate impact, demographic outcome equalization, or training-data bias amplification across protected classes. Same gap as EU AI Act Article 15.4.

Alternative: Fairness audit libraries (Fairlearn, AIF360); domain-specific impact assessment; demographic parity / equalized odds metrics.

4.6 MS-2.12 (environmental impact and sustainability)

styxx instruments operate at inference time on agent outputs; they do not measure training-energy consumption, inference-carbon footprint, or supply-chain environmental impact.

Alternative: ML CO2 calculators (CodeCarbon, MLCO2); cloud-provider sustainability reports; supply-chain LCA tooling.

5. Pre-registered falsification criteria

Same five criteria as the EU AI Act bridge ([papers/EU_AI_ACT_COMPLIANCE_2026.md](#) §5), adapted to the NIST context:

- **F1 (registry validity):** if any clause cites a Measure subcategory that does not exist in NIST AI 100-1, this paper is falsified.
- **F2 (calibrated-metric reproducibility):** if any cited AUC/accuracy number cannot be reproduced from the cited commit hash within ± 0.01 , that `PrimitiveCoverage` is falsified.
- **F3 (construct-ceiling discipline):** if a reviewer produces a published failure mode of a styxx primitive that is NOT mentioned in any `construct_ceiling` field, that entry is falsified.
- **F4 (boundary violation):** if a styxx primitive is later shown to produce evidence for an MS subcategory in `UNCOVERED_MEASURE`, the mapping was scoped too narrowly.
- **F5 (citation):** 1 independent citation (academic, federal procurement reference, state-level legislative citation, or commercial adoption note) by 2027-02-01 or methodology reassessed.

6. Reproducibility appendix

artifact	commit	path
this paper	this commit	<code>papers/NIST_AI_RMF_BRIDGE_2026.md</code>
NIST bridge module	this commit	<code>styxx/compliance/nist_ai_rmf.py</code>
shared dataclasses	this commit	<code>styxx/compliance/_common.py</code>
NIST tests (15)	this commit	<code>tests/test_compliance_nist_ai_rmf.py</code>
companion EU AI Act paper	<code>f10cab0</code>	<code>papers/EU_AI_ACT_COMPLIANCE_2026.md</code>
recursive-discipline paper v7	<code>cf14c83</code>	<code>papers/PAPER_recursive_discipline_2026_05_27.md</code>
MS-2.13 receipt (agent_audit instrument)	<code>3c24b5e</code>	<code>papers/agent-self-audit/FINDING_agent_claims.md</code>
MS-2.5 / MS-2.3 receipt (Baseline-019 PASS)	<code>17fdd97</code>	<code>submissions/baseline_019_openai_critique/scorecard.md</code>
MS-2.5 / MS-2.6 receipt (cognometric instruments)	<code>cf14c83</code>	<code>submissions/_results/leaderboard.json</code>

NIST AI 100-1 official source: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>. Measure-function subcategory text retrieved 2026-05-28 from <https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>.

7. Limitations + scope of applicability

1. **Measure function only.** Govern, Map, Manage are out of scope. A complete NIST AI RMF implementation requires all four functions.
2. **5 of 22 Measure subcategories covered.** MS-2.1, MS-2.2 (test set documentation, human-subject evaluation) and MS-1.x, MS-3.x, MS-4.x are out of scope for v0.1.
3. **Cross-jurisdictional, not cross-framework.** styxx primitives that satisfy NIST MS-2.5 may not satisfy EU AI Act Article 15.1(a) — the citation evidence may overlap but the contractual claims differ. Operators must conduct framework-specific conformity assessments.
4. **Voluntary regime.** NIST AI RMF is voluntary in US federal contexts. The bridge provides evidence; uptake depends on operator’s chosen compliance posture.

8. Conclusion

`styxx.compliance.nist_ai_rmf` v0.1 demonstrates that the EU AI Act bridge methodology transfers to a parallel jurisdictional regime with a different organizational structure. Both bridges share dataclasses, share the same kill-gate discipline, and reference the same underlying styxx primitives. v0.1 covers five MS-2.x Measure subcategories with five styxx primitives, enumerates six uncovered subcategories with alternative tool references, enforces structural integrity at test time, and pre-registers five falsification criteria.

Cross-jurisdictional AI deployments — increasingly common as US and EU operators serve the same downstream markets — benefit from a single open-source measurement primitive set that maps to both regimes. The `styxx.compliance` namespace establishes that pattern.

Not legal advice. Independent NIST AI RMF implementation review required for any production deployment.

Acknowledgments

Written 2026-05-28 alongside `styxx.compliance.nist_ai_rmf` v0.1 in the same session as the companion EU AI Act bridge. Subcategory text retrieved from the NIST AIRC web resource on the same day. The recursive-discipline methodology of `papers/PAPER_recursive_discipline_2026_05_27.md` v7 informs the pre-registration kill-gates and the construct-ceiling discipline.

This paper is offered to NIST AI RMF stakeholder communities, the AI Safety Institute Consortium (AISIC), federal procurement architects citing the RMF, state-level AI governance bill authors, and US enterprise compliance teams as an open-source measurement-methodology contribution. Comments, falsifications, and improvements are explicitly invited via the **fathom-lab/styxx** GitHub issue tracker.