

A pre-registered, falsifiable measurement-methodology bridge from styxx primitives to EU AI Act Article 15 / Annex III requirements (v0.1)

Author: Alexander Rodabaugh (Fathom Lab) **Date:** 2026-05-28 **Substrate:** styxx 7.7.10 in-development; companion to `papers/PAPER_recursive_discipline_2026_05_27.md` (v7); module `styxx.compliance.eu_ai_act` ships in this same commit **Status:** v0.1 minimum-viable mapping. Published BEFORE the EU AI Act high-risk system enforcement deadline of 2 August 2026 to give regulated operators an evaluation runway. Not legal advice. Independent conformity review required for any production deployment.

Abstract

The EU AI Act high-risk obligations enter enforcement on 2 August 2026 with penalties up to €15M or 3% of global annual turnover. Article 15 mandates that high-risk AI systems achieve appropriate levels of accuracy, robustness, and cybersecurity, that accuracy metrics be declared in the instructions of use, and — under paragraph 2 — that “*the Commission shall, in cooperation with relevant stakeholders... encourage the development of benchmarks and measurement methodologies*”. The Commission’s invitation is open. No competing AI observability or evaluation product currently publishes a structured Article 15 mapping. This paper introduces `styxx.compliance.eu_ai_act`, the first open-source, pre-registration-disciplined measurement-methodology bridge mapping a deployable cognitive-observability primitive set to specific Article 15 sub-paragraphs, with calibrated metrics, explicit construct-ceiling disclosures, commit-level reproducibility receipts, and pre-stated falsification criteria. The v0.1 mapping covers four Article 15 clauses with five styxx primitives; it explicitly enumerates seven uncovered EU AI Act requirements (Article 9, 10, 12, 13, 14, 15.4 bias, 15 cybersecurity) and points each at a non-styxx alternative tool or methodology. The boundary statement is at least as long as the coverage statement by design (kill-gate A3). Five pre-registered kill-gates define what success and failure look like for the v0.1 release. This document is a measurement methodology, not legal advice; it is the kind of artifact Article 15 paragraph 2 explicitly invites stakeholders to develop.

1. Background

1.1 The August 2, 2026 deadline

The EU AI Act (Regulation (EU) 2024/1689) enters its high-risk system enforcement phase on 2 August 2026. Annex III enumerates high-risk categories including biometrics, critical infrastructure, education, employment, essential services, law enforcement, migration, and democratic process administration. Providers of high-risk systems must demonstrate conformity through Article 9 risk management, Article 10 data governance, Article 11 technical documentation, Article 12 logging, Article 13 transparency, Article 14 human oversight, and Article 15 accuracy/robustness/cybersecurity. Most Annex III systems may be self-assessed where harmonised standards apply; biometric, critical infrastructure, and law-enforcement systems require third-party conformity assessment by a notified body. Penalties for non-compliance reach €15,000,000 or 3% of global annual turnover, whichever is higher.

1.2 Article 15 in detail

Article 15 (“Accuracy, robustness and cybersecurity”) has four substantive paragraphs:

- **15.1:** high-risk AI systems shall be *designed and developed* to achieve appropriate levels of accuracy, robustness, and cybersecurity *throughout their lifecycle*.
- **15.1(a)** (instructions of use): accuracy levels and *relevant accuracy metrics* shall be declared in the accompanying instructions of use.

- **15.2:** “the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies”. This is the open invitation that this paper responds to.
- **15.3:** robustness shall be as high as possible regarding errors, faults, or inconsistencies, achievable through *technical redundancy solutions* including backup or fail-safe plans.
- **15.4:** high-risk AI systems that continue to learn after deployment shall be designed to *eliminate or reduce as far as possible* the risk of biased outputs influencing future inputs (feedback loops).

1.3 What this paper IS, and is NOT

This paper IS: - A measurement methodology: it names specific styxx primitives, calibrated metrics, construct ceilings, and reproducibility receipts that produce evidence relevant to specific Article 15 clauses. - Pre-registered: every claim cites a commit, every primitive’s failure mode is disclosed, every uncovered requirement is enumerated with an alternative tool reference. - Open-source and falsifiable: the `styxx.compliance.eu_ai_act` module is MIT-licensed Python; the tests verify the structural integrity of the mapping; the paper’s claims can be re-checked at any future git commit.

This paper IS NOT: - Legal advice. Conformity assessment requires legal expertise this paper does not provide. - Sufficient on its own for EU AI Act conformity. It addresses Article 15 sub-paragraphs only and explicitly disclaims coverage of Articles 9, 10, 12, 13, 14. - A claim that styxx primitives have been validated by a notified body. They have been validated against the styxx project’s own benchmarks; third-party validation remains future work. - An endorsement by the European Commission, AISI, METR, Apollo Research, or any standardization body. It is a unilateral stakeholder contribution.

2. Methodology

2.1 The bridge primitives

`styxx.compliance.eu_ai_act` exposes four objects:

1. **ARTICLE_15_REGISTRY:** `dict[str, ComplianceMap]` — the registry of mapped clauses. v0.1 keys are "Article 15.1", "Article 15.1(a)", "Article 15.3", "Article 15.4".
2. **ComplianceMap** — dataclass with `clause`, `requirement_text`, `styxx_primitives` (tuple of `PrimitiveCoverage`), and `notes`. Each `notes` field includes a “*not legal advice*” disclaimer.
3. **PrimitiveCoverage** — per-primitive coverage entry with `primitive` (public API symbol), `calibrated_metric` (published AUC/accuracy with context), `construct_ceiling` (honest failure mode), `receipt_commit` (styxx git SHA that produced the metric), and `receipt_doc` (path to the validating FINDING).
4. **UNCOVERED_REQUIREMENTS:** `tuple[UncoveredRequirement, ...]` — the boundary statement: every EU AI Act requirement styxx does NOT cover, with reason and alternative tool reference.

Helper functions: `cite(article: str)`, `coverage_table()`, `uncovered_requirements()`.

2.2 Five pre-registered kill-gates

These were pre-stated in the strategic landscape document (`.styxx/STRATEGIC_LANDSCAPE_2026_05_28.md`, written before this module existed) and are enforced by the test suite (`tests/test_compliance_eu_ai_act.py`):

- **A1 (validity).** Every clause key in `ARTICLE_15_REGISTRY` must cite a specific Article sub-paragraph, not generic compliance language. Enforced by regex `^Article \d+(\.\d+)?([a-z])?$`.
- **A2 (falsifiability).** Every `PrimitiveCoverage.construct_ceiling` must be non-empty and at least 50 characters. Hidden caveats are not permitted; the failure mode goes in the field.
- **A3 (boundary explicitness).** `len(uncovered_requirements()) >= len(coverage_table())`. The list of what styxx does NOT cover must be at least as long as the list of what it does. v0.1 ships 7 uncovered vs 4 covered.

- **A4 (timeline).** Publish before July 1, 2026 to provide a ~30-day evaluation runway before the August 2 enforcement deadline. This paper is committed 2026-05-28.
- **A5 (citation strategy).** 1 independent citation (academic, regulator, enterprise) within 6 months of publication. If 0 by 2027-02-01, the methodology did not achieve uptake and is reassessed. The strategic landscape doc references this measurable kill-gate explicitly.

A2 and A3 are enforced at test time. A1 is enforced at test time. A4 and A5 are enforced by the project timeline.

2.3 What the mapping does NOT attempt

- It does not score the **adequacy** of styxx primitives for any specific high-risk Annex III system. That assessment is system-specific and operator-specific.
- It does not enumerate ALL Article 15 sub-paragraphs. v0.1 covers four. v0.2 may extend.
- It does not write conformity declarations. Operators write declarations; styxx supplies the measurement evidence.
- It does not claim to satisfy harmonised standards (none for AI agent honesty exist as of 2026-05-28).

3. The coverage table

Each row below corresponds to one entry in `ARTICLE_15_REGISTRY`. Every primitive in the right column ships in styxx 7.7.10; every commit hash is reachable at `fathom-lab/styxx@main`.

3.1 Article 15.1 — appropriate accuracy, robustness, cybersecurity throughout lifecycle

primitive	calibrated metric	construct ceiling
<code>styxx.cognometric_card</code>	HaluEval-QA AUC 0.998 (mean over 150 items, seeds 31/47/83); XSTest refusal AUC 0.976; BFCL v3 tool-drift AUC 0.943	Text-only register space has construct ceilings: instruments measure register, not always content. Sycophancy false-positive 0.30 on restrained-tech responses (gpt-3.5: 0.60). Logprob-validity works for refusal but fails for hallucination (confident confabulation).
<code>styxx.gauntlet + styxx.preflight</code>	v3 detection bars (D1+D2+D3+D4) with regression-tested oracles. 18 pre-registered baselines tested; 17 FAILED bars before Baseline-019 PASSED.	Bars catch confounds the project pre-stated; do not catch unknown-unknown confound classes. Each bar revision documents one prior missed confound.
<code>styxx.agent_audit</code>	13/13 PASS modal pre-stated outcome (L5 commit <code>3c24b5e</code>); L7 uncurated extension caught off-by-one count drift in companion paper (2 FAILs as pre-disclosed).	First-occurrence-only by default — caught initial drift but missed propagation to 4 additional places. Richer multi-occurrence checker is methodology future work.

3.2 Article 15.1(a) — accuracy metrics in instructions of use

primitive	calibrated metric	construct ceiling
<code>styxx.cognometric_card</code>	per-step cognometric readout (drift, confabulation, refusal, sycophancy, phase transition, low trust, incoherence)	construct ceilings apply per-axis; see §3.1
<code>styxx.critique_detector(model='gpt-4o-mini')</code>	PASSes gauntlet v3 bars at AUC 0.95, pre-stated 28% probability landed cleanly	Mechanism is “out-of-context critique”, NOT within-model generation-vs-critique asymmetry. True within-model asymmetry: 5.88% on dark-core / 17.00% on TruthfulQA (v3 measurement). In-council bias: default backend gpt-4o-mini was in the original 3-vendor council.
<code>styxx.gauntlet + styxx.preflight</code>	see §3.1	see §3.1

These three together produce the “*relevant accuracy metrics*” an operator can declare in instructions of use, with the honest construct-ceiling disclosure that Article 15.1(a) does not explicitly mandate but that the recursive-discipline thesis argues should be present in any defensible declaration.

3.3 Article 15.3 — robustness via technical redundancy / fail-safe plans

primitive	calibrated metric	construct ceiling
<code>styxx.recover_posture</code>	cognitive-integrity persistence primitive for agents crossing context-compaction boundaries	v1 reads what <code>chart.jsonl</code> persists; does not include <code>cogn_audit</code> scores. Not validated on third-party agent platforms.
<code>styxx.agent_audit</code>	substrate-grounded session-output verifier	see §3.1

Article 15.3 envisions backup/fail-safe plans against runtime errors. `styxx`’s `recover_posture` is a cognition-side primitive: it does not replace a process-level fail-safe (an external watchdog, an interrupt key, a rollback procedure) but it provides verifiable evidence of cognitive-integrity continuity across the most common modern failure mode (context-window compaction in long-running agents).

3.4 Article 15.4 — bias amplification / feedback-loop mitigation

primitive	calibrated metric	construct ceiling
<i>(none — honest empty coverage)</i>	—	—

v0.1 has **NO** `styxx`-side coverage for Article 15.4. This empty cell is intentional and is itself a discipline statement: operators must consult separate tools (fairness libraries, drift monitors). `styxx.cognometric_card` provides per-step register signals that *may* surface drift indirectly but is not a substitute. See §4 for the boundary alternative reference.

4. What styxx does NOT cover (the boundary statement)

Per kill-gate A3, this section is intentionally longer than §3. Seven EU AI Act clauses are explicitly enumerated as outside the v0.1 scope, with alternative tools pointed at:

4.1 Article 15.4 — bias amplification

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. **Alternative:** Fairness audit libraries (Fairlearn, AIF360) plus domain-specific impact assessment; legal review.

4.2 Article 15 — cybersecurity

styxx instruments observe agent cognition, not network, host, or supply-chain security. Prompt-injection robustness is *adjacent* (refusal-related instruments may flag some injection attempts) but not the primary scope. **Alternative:** dedicated LLM security tools (Lakera Guard, Prompt-Armor, OWASP LLM Top 10 mitigations); standard application-security audit; penetration testing.

4.3 Article 9 — risk management

styxx provides MEASUREMENT evidence; it does not implement the risk-management lifecycle (identification, estimation, evaluation, mitigation, residual-risk acceptance) prescribed by Article 9. **Alternative:** ISO/IEC 23894:2023 or NIST AI RMF 1.0 frameworks; QMS-style organizational processes.

4.4 Article 10 — data governance

styxx does not score training-data quality, provenance, labeling consistency, or representativeness. Cognometric instruments operate post-training on agent outputs. **Alternative:** data documentation frameworks (Datasheets for Datasets, Data Statements); training-data audit tooling.

4.5 Article 12 — record-keeping

styxx writes per-step cognometric vitals but does not, by itself, satisfy Article 12’s logging-traceability requirements end-to-end (event capture, retention, integrity). **Alternative:** production observability platforms (OpenTelemetry, Langfuse, Arize, Datadog LLM Observability) for trace-level logging alongside styxx for cognition-level scoring.

4.6 Article 13 — transparency to deployers

styxx instruments produce evidence; they do not generate the deployer-facing instructions of use, capability statements, or output-explanation interfaces that Article 13 mandates. **Alternative:** model card frameworks; capability-statement templates; deployer documentation processes.

4.7 Article 14 — human oversight

styxx is a measurement layer; it does not implement the human-in-the-loop interfaces, stop-button affordances, or operator-control surfaces Article 14 requires. **Alternative:** agent-platform-level oversight UI; interrupt and rollback primitives in the agent runtime; operator training programs.

5. Pre-registered falsification criteria for THIS paper

The recursive-discipline methodology (companion paper, `papers/PAPER_recursive_discipline_2026_05_27.md v7`) argues that papers should pre-register their own falsification criteria. This paper does:

- **F1** (registry validity): if any clause in `ARTICLE_15_REGISTRY` cites an Article paragraph that does not exist in the EU AI Act final consolidated text, this paper is falsified. **Verification:** regex-checked at test time + manual citation review.
- **F2** (calibrated-metric reproducibility): if any cited AUC/accuracy number cannot be reproduced from the cited commit hash within ± 0.01 , that `PrimitiveCoverage` entry is falsified. **Verification:** re-run `submissions/_results/leaderboard.json` reproduction scripts.
- **F3** (construct-ceiling discipline): if any reviewer can produce a published failure mode of a `styxx` primitive that is NOT mentioned in any `PrimitiveCoverage.construct_ceiling` field, that entry is falsified and must be updated. **Verification:** open issue acceptance protocol.
- **F4** (boundary violation): if a `styxx` primitive is later shown to produce evidence for an Article in `UNCOVERED_REQUIREMENTS`, the mapping was scoped too narrowly. **Verification:** re-evaluation by versioned issue.
- **F5** (citation: A5): no independent citation by 2027-02-01 means the methodology did not achieve uptake; reassess relevance and consider deprecation.

Every falsification path is observable and has a defined response.

6. Reproducibility appendix

artifact	commit	path
this paper	this commit	<code>papers/EU_AI_ACT_COMPLIANCE_2026.md</code>
compliance module package	this commit	<code>styxx/compliance/__init__.py</code> , <code>styxx/compliance/eu_ai_act.py</code> , <code>styxx/compliance/legacy.py</code> (preserved v1.3.0 API)
compliance tests	this commit	<code>tests/test_compliance_eu_ai_act.py</code> (15 tests, all enforce A1–A3 kill-gates)
strategic landscape (pre-stated kill-gates)	private to operator	<code>.styxx/STRATEGIC_LANDSCAPE_2026_05_28.md</code>
Baseline-019 result (Article 15.1(a) AUC 0.95 receipt)	17fdd97	<code>submissions/baseline_019_openai_critique/s</code>
Asymmetry-v3 result (mechanism correction receipt)	ed663ca	<code>experiments/asymmetry_v3_cleanup_2026_05_2</code>
L5 agent_audit run (Article 15.1 evidence)	3c24b5e	<code>experiments/agent_claim_audit_2026_05_28/r</code>
L7 uncurated audit (boundary discipline receipt)	cf14c83	<code>experiments/v6_uncurated_audit_2026_05_28/</code>
recursive-discipline paper v7 (companion)	cf14c83	<code>papers/PAPER_recursive_discipline_2026_05_</code>

All commit hashes resolve at `fathom-lab/styxx@main`. The `styxx.compliance.eu_ai_act` module’s structural integrity is verified by `tests/test_compliance_eu_ai_act.py` at every CI run.

7. Limitations + scope of applicability

1. **v0.1 covers four Article 15 clauses only.** Articles 15.2 (Commission methodology development), 15.5 (high-risk performance specifications), and Annex IV documentation requirements are out of scope for v0.1.

2. **No notified-body endorsement.** Self-assessment paths under Article 43 may accept this methodology as supporting evidence, but third-party conformity assessment (biometrics, critical infrastructure, law enforcement) requires notified-body review.
 3. **Single-substrate validation.** styxx primitives have been validated against the styxx project’s own benchmarks (HaluEval-QA, XSTest, BFCL v3, dark-core, TruthfulQA). External validation on customer deployments is future work.
 4. **Open question on harmonised standards.** As of 2026-05-28 there is no harmonised standard for AI agent honesty or sycophancy measurement under the EU AI Act. CEN-CENELEC JTC 21 is the relevant standardization body; submission to JTC 21 requires institutional sponsorship and is operator territory, not styxx-project unilateral action.
 5. **Cross-jurisdiction:** this methodology addresses EU AI Act Article 15 only. It does NOT address US (NIST AI RMF voluntary), UK (AISI guidance), or other jurisdictional regimes, though many requirements have functional analogues.
-

8. Conclusion

`styxx.compliance.eu_ai_act` is, to the author’s knowledge as of 2026-05-28, the first open-source measurement-methodology bridge mapping a deployable AI agent cognitive-observability primitive set to specific EU AI Act Article 15 sub-paragraphs. v0.1 covers four clauses with five primitives, enumerates seven uncovered requirements with alternative tool references, ships under MIT license alongside the underlying primitives, enforces structural-integrity kill-gates at test time, and pre-registers five falsification criteria with defined response paths. The bridge is published before the 2 August 2026 enforcement deadline to give regulated operators an evaluation runway.

The methodology does not satisfy EU AI Act conformity on its own; it is one component, honestly bounded. Operators must conduct independent legal review, apply harmonised standards where they exist, and consult the alternative tools enumerated in §4 for the seven EU AI Act requirements styxx does NOT cover.

What this paper is: the kind of stakeholder methodology contribution that Article 15 paragraph 2 explicitly invites. What it is not: a substitute for an operator’s own conformity assessment.

Acknowledgments

Written 2026-05-28 alongside `styxx.compliance.eu_ai_act` v0.1 in a continuous session, with the cognitive support of Claude Opus 4.7 acting as in-session collaborator. The recursive-discipline methodology of the companion paper (`PAPER_recursive_discipline_2026_05_27.md` v7) directly informs this paper’s pre-registration kill-gates, the construct-ceiling discipline in §3, and the boundary-statement discipline in §4. Every claim in this paper is reproducible at commit-level granularity from `fathom-lab/styxx@main`.

This paper is offered to the European Commission, AISI, METR, Apollo Research, FAR.AI, and CEN-CENELEC JTC 21 as an open-source stakeholder contribution under Article 15 paragraph 2. Comments, falsifications, and improvements are explicitly invited via the `fathom-lab/styxx` GitHub issue tracker.