

TierGuard: A Deterministic Triage-and-Abstention Baseline for Article 5, Annex III and Article 50 of the EU AI Act

Rashad Kamsheh
Independent Researcher
Sweden

rashad.kamsheh@outlook.com

Abstract—Under the EU AI Act (Regulation (EU) 2024/1689), an AI system’s risk tier is the load-bearing first determination for many system-level obligations, yet making that determination defensibly at inventory scale remains an open, operational problem. We present TierGuard, a *lexical, deterministic triage-and-abstention baseline* for the Act’s Article 5 prohibited practices, Annex III high-risk use-cases, and Article 50 transparency duties; Annex I routing is out of scope and flagged as not assessed, and of Article 50 only 50(1), 50(2) and the deep-fake limb of 50(4) are encoded. From a free-text system description the engine either issues a candidate tier bound to its governing provision, or returns an INDETERMINATE, review-required outcome carrying candidate provisions and missing facts, deliberately separating workflow escalation from legal conclusion. Every decision is a machine-readable, version-pinned record preserving concurrent Article 50 flags. The evaluation is safety-oriented and abstention-aware: on the primary post-freeze set (n=45) the baseline reaches 58% correct-or-referred with 40% silent under-classification (95% CI 27-55%), 27% severe, i.e. missed high-risk or prohibited references, versus 71% (56% severe) for a naive internal keyword baseline; on 22 interpretive labels derived from the Commission’s draft high-risk guidelines, 91% (20/22). Two engine-blind annotators (n=57, Cohen’s kappa 0.971) reproduce the author’s tier labels at 96%, leaving the engine’s figures unchanged: a practitioner reproducibility check, not legal validation. Compliance-washing and multilingual probes expose the lexical recall limits: off English, silent under-classification rises to 58-83%. These figures establish a measured floor and an evaluation protocol for free-text triage, not a deployable screen; the public artifact and benchmark suite are released for semantic and LLM-based approaches to beat.

Index Terms—EU AI Act, risk classification, triage, abstention, RegTech, compliance-by-design, design science research.

I. INTRODUCTION

Regulation (EU) 2024/1689 (the AI Act) governs how AI systems may be built, placed on the market, and used across the Union, applying *inter alia* to providers placing systems on the Union market or into service there, and to providers and deployers where the system’s output is used in the Union (Art. 2) [1]. Prohibitions apply since February 2025. The Digital Omnibus on AI was published as Regulation (EU) 2026/1744 on 24 July 2026 and entered into force on 27 July 2026 [2]; it defers Annex III high-risk obligations to 2 December 2027 and Annex I product-embedded systems to 2 August 2028,

and inserts two new prohibitions into Art. 5(1), points (ba) and (bb), which apply from 2 December 2026. Many system-level obligations turn on a prior determination: what risk tier is this AI system? Misclassify, and the downstream controls are misdirected.

Today that determination is largely manual. The Commission’s beta Compliance Checker [3] offers questionnaire-style guidance on which rules may apply, but the determination itself remains a per-system human judgement that is slow, varies between reviewers, and rarely leaves a per-decision audit trail tied to the governing provision. At inventory scale of hundreds of systems per organisation, a first-pass triage with an explicit audit trail is the missing operational layer.

We contribute TierGuard, a lexical, deterministic triage-and-abstention baseline for three of the Act’s provision families. We are explicit about what it is: pattern-based detection with synonym normalisation over free-text descriptions: not semantic understanding and not an impact calculus; the Act’s impact judgements live in the rules’ citations, and its recall limits are measured, not hidden. Rule-based classification, abstention and human review are each established; what is new is their combination for this determination, and its measurement. The contributions are: (C1) an outcome model in which uncertainty is a first-class INDETERMINATE outcome routed to review, not a risk tier; (C2) a versioned rule set in which every candidate decision is bound to the provision it cites; (C3) an abstention-aware evaluation protocol that scores error direction, silent under-classification rather than accuracy, on post-freeze, regulator-derived, adversarial and multilingual sets, released as a public benchmark suite; and (C4) a machine-readable, version-pinned decision record preserving concurrent Article 50 obligations. We claim no more: the engine is a measured floor for the semantic and probabilistic strands to beat, and the design principles (cite every decision, fail safe to escalation, keep uncertainty first-class) are a design conjecture, not a validated theory.

II. BACKGROUND AND RELATED WORK

A. The Act’s tier structure

A practice satisfying Art. 5’s elements and falling outside its exceptions is *prohibited*. A system is *high-risk* under

Art. 6(1) (safety components of Annex I products) or Art. 6(2) with Annex III’s eight use-case domains; a narrow Art. 6(3) derogation requires a listed condition, is unavailable where natural persons are profiled, and must be documented. Art. 50 transparency duties are not exclusive: per Art. 50(6) they continue to apply where a system is also high-risk. All other in-scope systems are treated as minimal for this triage, although orthogonal obligations may remain. High-risk classification activates the Art. 9-15 stack plus conformity assessment, CE marking and EU-database registration.

B. Related work

Three strands automate parts of this determination. *Manual frameworks*: Schuster et al. [4] provide a four-step checklist for GPAI systems, validated qualitatively. *Probabilistic systems*: LLM-based reasoning supports compliance and robustness argumentation [5], and RAG systems have been evaluated on risk-level classification, reaching 0.87/0.85 F1 for prohibited/high-risk scenarios on an open benchmark [6], but F1, a precision-recall summary, reveals neither how many high-risk systems are missed nor whether misses fail silent; deterministic verification work [7] begins after the tier is known. *Semantic-web frameworks*: closest to our aims, the AIRO ontology represents AI-risk information for high-risk documentation [8]; VAIR formalises Annex III concepts for identifying high-risk applications [9]; and a recent framework provides SHACL/N3 automated rule-checking of risk categories, machine-readable intended-purpose policies, and compliance queries over RDF-described systems [10]; adjacent work makes AI compliance evidence itself machine-readable [11]. Table I positions these. The semantic strand presumes a system already modelled in structured vocabularies and reports no benchmarked accuracy, abstention behaviour, or adversarial robustness; the probabilistic strand takes free text but offers no reproducibility or bounded silent error. Closest are a decision-tree classifier evaluated quantitatively against the Act [12] and an artifact reporting automated risk assessment at 88% accuracy [13]; both take structured or questionnaire input and report accuracy without an abstention path or an error-direction breakdown, so a silent miss and an over-tier count alike. TierGuard occupies an underexplored combination rather than an empty cell: an *evaluated*, free-text, deterministic triage baseline with abstention and error-direction testing: a baseline that semantic and LLM approaches can and should be measured against, and with which they compose (a TierGuard escalation can seed an AIRO/VAIR-style structured assessment).

III. METHODOLOGY

A. Outcome model: four outcomes plus concurrent Article 50 flags

From a description, the engine returns one tier-track outcome (Fig. 1): *prohibited-candidate* (an Art. 5 practice signal; gated for legal confirmation, since the prohibitions carry element tests and express exceptions which review must

apply); *high-risk-candidate* (a specific Annex III use-case matched; gated for confirmation and the Art. 6(3) derogation check, a derogation that holds only where the system poses no significant risk of harm, including by not materially influencing the outcome of decision-making, and which under Art. 49(2) still requires EU-database registration); *minimal*, always qualified “Annex I route not assessed”; or *INDETERMINATE_REVIEW_REQUIRED*, together with **zero or more concurrent Article 50 transparency flags**, recorded on every outcome per Art. 50(6): per-paragraph, actor-specific duties, not a mutually exclusive class. A record whose only trigger is an Art. 50 flag is reported as the evaluation’s “limited” row. The indeterminate outcome is issued when a regulated context or a consequential automated decision about persons is detected without a resolved use-case; its record carries candidate provisions and missing facts (intended purpose, use-case, Art. 6(3) conditions). This separation is deliberate: legally, uncertainty is not an Art. 6 route to high-risk status; operationally, it must still never pass silently. All candidate outcomes are candidates pending human confirmation.

B. Lexical detection, candidly

Detection is pattern matching: a versioned rule set (10 Art. 5 rules, 8 Annex III use-case rules and one weak regulated-context rule, and 3 Art. 50 rules covering Arts. 50(1), 50(2), and the deep-fake limb of 50(4), with the responsible actor recorded. Art. 50(3) and 50(4)’s public-interest-text limb are not yet represented; each rule carries citation and confidence) over normalised text, plus two deterministic hardening components: synonym normalisation, and a consequential-decision-about-persons net that triggers the indeterminate outcome. The engine is surface-form dependent by construction; the compliance-washing and multilingual results (Section IV) measure exactly that dependence. Structured intake (intended purpose, operator role, territorial scope, affected persons, evidence provenance) is the identified next layer: a free-text description is sufficient for triage, not for a legal conclusion, which is why conclusions are never issued without review.

C. The decision record

Every outcome is emitted as a machine-readable passport: outcome; citation; rules fired; confidence; gate flag and kind (confirmation vs. review); concurrent Art. 50 flags (per Art. 50(6)); GPAI status as an orthogonal, model-level flag (Chapter V); for indeterminate outcomes, candidate provisions and missing facts; a deterministic counterfactual; the Annex-I not-assessed qualifier; and provenance (rule-set version, Act version). Because the Annexes can be amended, version pinning supports a regulatory diff: re-running an inventory under a new rule-set version reports which systems change outcome, and the Omnibus’s newly adopted Art. 5 prohibition on non-consensual intimate imagery [2] concretely illustrates the use.

TABLE I
POSITIONING AGAINST PRIOR APPROACHES TO AI ACT RISK DETERMINATION. TIERGUARD OCCUPIES THE EVALUATED, FREE-TEXT, ABSTENTION-AWARE CELL.

Approach	Input	Scope	Referral / evaluation
Checklist [4]	human Q&A	GPAI tiers	human throughout; 3 cases, qualitative
LLM / RAG [5], [6]	free text	risk levels	none reported; F1 only, no error direction
AIRO / VAIR [8], [9]	RDF model	Annex III docs	n/a (a representation); no accuracy benchmark
Semantic fw. [10]	RDF model	risk, docs, reg.	rule-check verdicts; no accuracy / adversarial eval
Checker [3]	web form	applicable rules	human throughout; n/a
TierGuard	free text	Art. 5 + Annex III + Art. 50	gate + INDETERMINATE; post-freeze, adversarial, multilingual, error-direction

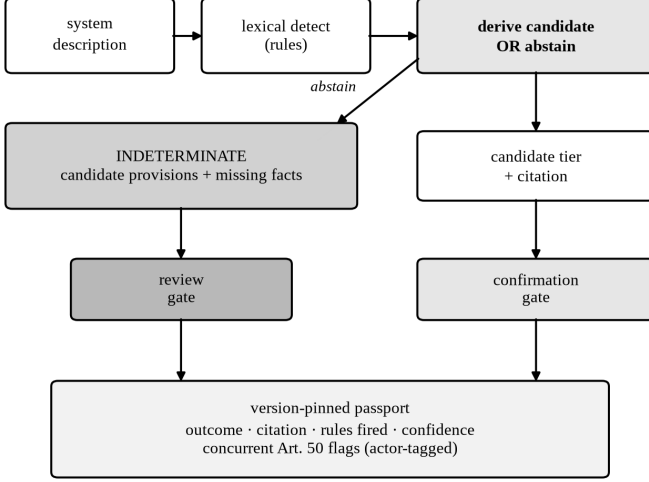


Fig. 1. Pipeline: lexical detect, derive-or-abstain, ground. Candidate tiers carry an Article/Annex citation; indeterminate outcomes carry candidate provisions and missing facts; every record is version-pinned and preserves concurrent Art. 50 flags.

D. Evaluation design

The artifact and its evaluation follow design science research practice [14], [15], with the evaluation route chosen for risk and effectiveness [16]. We evaluate by error direction and abstention, not accuracy alone. The headline failure metric is *silent under-classification*, a silent duty miss: a required prohibited- or high-risk-candidate outcome, or a required Article 50 flag, that is neither surfaced nor gated. Its severe subset, *silent missed-risk*, a high-risk or prohibited reference so missed, is reported alongside, since a missed Art. 50 duty and a missed prohibition are not the same harm. The safe-handling rate (correct or escalated) is always paired with the gate rate, since escalating everything would trivially maximise it; coverage (share decided rather than abstained) is reported alongside. Sets: a development-influenced evaluation set (n=86; Cycle 2’s mechanisms were designed from development failures and aggregate Cycle 1 metrics, so it is not strictly held out); the **primary post-freeze set**, 46 cases authored after the rule set was frozen; 13 worked examples from the Commission’s *adopted* guidelines on prohibited practices

[17]; a 22-item post-freeze set from its *draft* high-risk classification guidelines, used as *guideline-derived interpretive labels*, since that draft is non-binding and its consultation closed 23 July 2026, final guidelines pending [18]; 14 real, named systems from public documentation (author-selected and author-labelled, small); a 20-item adversarial compliance-washing set; and an EN/DE/FR/AR multilingual slice (author-translated). Reference labels are author-assigned from the Act and fixed before scoring; a two-annotator reproducibility check of the tier labels is reported in Section IV-D. A SHA-256 manifest fixes the package contents, though the released archive carries no repository history, so the freeze is an author assertion the manifest supports rather than proves; the post-freeze tests were evaluated once.

IV. RESULTS

A. Primary result: post-freeze triage quality

Table II reports all sets under the abstention-aware model. Annex I product-safety cases are excluded throughout (three in the evaluation corpus, two in the real-system set); the engine does not attempt that route and flags it on each record, so scoring them measured a determination it declines to make. On the primary post-freeze set the lexical baseline reaches 58% correct-or-referred (26/45; Cycle 1: 53%) with 40% silent under-classification (18/45; Wilson 95% CI 27-55%) at a 29% gate rate and 73% coverage. The severe subset: of the 25 high-risk-or-prohibited references, 13 (52%) are safely handled and 12 are silently missed, i.e. **severe-risk silent misses 12/45 = 27% overall, 12/25 = 48% class-conditional**; the remaining 6 silent under-classifications are Art. 50 reference cases. Under the same definitions a *naive internal keyword baseline*, a floor we wrote ourselves rather than a published system, is 71% silent (32/45) and 56% severe here and 67% silent on the n=86 set. Exact tier agreement on decided cases is 42% (14/33); macro-F1 over decided cases is 0.25, low and reported deliberately: most residual misses are semantic paraphrases that lexical rules cannot reach, the measured boundary of this baseline. On the 22 guideline-derived interpretive labels, correct-or-referred is 91% (20/22) with 2 silent at a 68% gate rate; on the 13 prohibited-practices examples, 77% (10/13). On 12 in-scope real systems: 12/12, 0 silent. Cost is of the order of 100 µs per description, remeasured per run.

TABLE II

ABSTENTION-AWARE RESULTS PER SET, ANNEX I PRODUCT-SAFETY CASES EXCLUDED THROUGHOUT. C-O-R = CORRECT-OR-REFERRED, I.E. CORRECT OR ESCALATED TO REVIEW: IT MEASURES ROUTING, NOT SUCCESSFUL RESOLUTION, SINCE THE HUMAN STAGE IS UNEVALUATED. SILENT = UNGATED UNDER-CLASSIFICATION, ANY TIER (FOR THE BINARY DRAFT-GUIDELINE SET: A MISSED HIGH VERDICT); COV. = DECIDED, NOT ABSTAINED.

Set (n)	C-o-R	Silent	Gate	Cov.	Exact/n
Post-freeze 45, C2 (prim.)	26 (58%)	18 (40%)	29%	73%	14 (31%)
Post-freeze 45, C1	24 (53%)	20 (44%)	24%	78%	14 (31%)
Evaluation 86, C2 (dev-infl.)	54 (63%)	31 (36%)	35%	70%	28 (33%)
Naive keyword baseline (86)	28 (33%)	58 (67%)	0%	100%	28 (33%)
Draft-guideline labels (22)	20 (91%)	2 (9%)	68%	50%	8 (36%)
Prohib.-pract. (13)	10 (77%)	3 (23%)	77%	54%	3 (23%)
Real systems (12, in scope)	12 (100%)	0 (0%)	50%	67%	8 (67%)
Adv. washing (20), C2	12 (60%)	8 (40%)	60%	50%	n/a

B. Hardening ablation, adversarial and multilingual behaviour

Cycle 2's two components (synonym normalisation + persons-net) cut adversarial compliance-washing slip-through from 60% to 40% (silent, n=20) and post-freeze silent under-classification from 44% to 40% (Fig. 2). The post-freeze change is modest and not statistically distinguishable from chance: 2 cases improved, none worsened, exact McNemar $p=0.50$ on the paired comparison, and the two intervals overlap heavily (Wilson 95% CI 31-59% versus 27-55%). We report it as a direction, not a demonstrated effect, converting misses into gated or indeterminate outcomes rather than silent errors. The washing set shows the lexical dependence plainly: euphemised descriptions of consequential systems evade specific rules, and the persons-net recovers only those that still mention a decision about people. Off English, exact agreement collapses: on 12 systems each rendered in four languages (12 cases per language, faithful manual translations), exact agreement falls from 92% in English to 25%/25%/17% in DE/FR/AR, with silent under-classification rising from 0% to 67%/58%/83%; the cause is mechanical, since English patterns do not match non-English tokens, so a translation front-end or per-language rules are required layers, not options.

C. Upper bound (not an evaluated system)

If the optional, always-gated semantic detector of the architecture recovered every residual use-case, silent under-classification would fall to 0% by construction. We state this perfect-residual upper bound only to size the residual; it is an analytical ceiling, not a result, and a live detector stage with fixed prompts and variance reporting is future work.

D. Reference-label reproducibility

Two annotators, a data engineer and a data scientist, neither a legal practitioner, independently labelled 57 system descriptions (tier, provision, Article 50 set) from a written guide, blind to the engine and to each other; agreement was computed before adjudication and the engine run only after the agreed labels were fixed. Tier agreement was 56/57 (Cohen's kappa 0.971), and wherever both named a provision they named the same one (28/28). Against the 52 agreed units

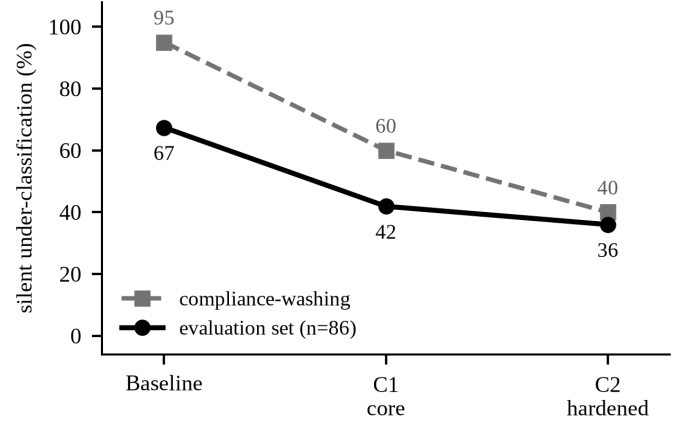


Fig. 2. Silent under-classification across the two hardening cycles on the evaluation (n=86) and adversarial compliance-washing sets. The perfect-residual upper bound is excluded from this measured trend (see IV-C).

the author-assigned tier labels reproduce at 50/52 (96.2%; Wilson 95% CI 87-99%), the two exceptions being cases the author tiered stricter (high) than the annotators (minimal). Scored by one rule on the same 52 units, the engine is 16/52 (30.8%) silent and 35/52 (67.3%) safely handled under the author's labels, and identically under the agreed labels: whose labels are used changes nothing about how the engine scores. Five rows are unresolved and deliberately not adjudicated; four turn on whether Article 50 duties survive an Article 5 prohibition, a split the Regulation does not settle. Two limits bound this check: the annotators are practitioners, so agreement establishes reproducibility, not legal correctness; and the guide as issued carried an Article 6(3) instruction error that could only have pushed labels towards high, so the figures are round-1 estimates exposed on the 19 units both labelled high, re-run first under the corrected guide (Section IV-E).

E. Threats to validity and validation plan

Reference labels are author-assigned, with the reproducibility check above as the only independent evidence to date; the draft-guideline labels inherit that draft's non-binding, interpretive status (the prohibited-practices examples come

from adopted, though still interpretive, guidance); the real-system set is small (14 systems) and author-selected; the evaluation set is development-influenced (hence the post-freeze primary); the engine is free of run-to-run variance while benchmark-sampling uncertainty remains (silent under-classification on the evaluation set: 36.0%, 95% CI 26-47%). The comparator is a limit too: the keyword baseline is a floor we wrote, because no published free-text classifier was runnable on this benchmark at freeze and the semantic-web systems take RDF input the corpora lack; the suite is released so that comparison can be made on equal terms. The independent check is specified, and its protocol ships in the study package ahead of any expert data: 5-8 AI-law or compliance practitioners, including legally qualified raters, engine-blind and mutually blind, re-label a stratified sample of about 120 descriptions (all tiers, Art. 6(3) and Art. 5 edges, washing items, real model cards) under the corrected guide, citing the provision; Fleiss or Krippendorff agreement is reported before adjudication, adjudicated labels replace ours, and three pre-stated hypotheses are tested: expert agreement is high but below 1.0; disagreement concentrates on gated items (permutation test); and on non-gated items engine-majority kappa is non-inferior to mean pairwise expert kappa, margin fixed at registration. The round also re-runs the 19 Art. 6(3)-exposed units and every Article 50 label, and a second post-freeze wave of at least 100 cases sampled from EU database registrations and public incident inventories, not authored, narrows the intervals reported here.

V. DISCUSSION

A. Implications

For providers, TierGuard is a first-pass triage: a cited candidate outcome or an explicit review-required record in approximately 100 μ s in the author's recorded environment, with a context-dependent share escalated (29% on the post-freeze set, 68% on guideline-derived labels). For reviewers, indeterminate records arrive with candidate provisions and missing facts, so review is directed, not from scratch. What the primary result licenses is narrow: at 40% silent under-classification, four in ten post-freeze cases carrying a duty, and 12 of 25 high-risk or prohibited references, would pass an unassisted deployment without a flag. As a stand-alone screen that is not acceptable for a safety-oriented control, and no such use is claimed; the figure is a floor (what English lexical rules over free text reach, every residual miss catalogued in the artifact) and a direction (the abstention path turns further misses into review rather than errors). The deployment sense is a first pass in front of human review, never in place of it, with gate rate and silent rate monitored as paired service levels. For the semantic-web strand [8]–[10], the baseline is complementary: escalations identify which systems merit full structured modelling, and our benchmark suite offers the evaluation layer that strand has not yet reported. The audit-supporting record (citation, rules, versions, counterfactual) is designed for challenge and replay; whether it satisfies

specific audit criteria is untested. An illustrative weighted decision-cost analysis (weights proxying the Art. 99 penalty asymmetry; full sweep in the artifact's cost.py) favours the gated engine over the silent baseline across the weightings we tested; we treat it as a scenario analysis only. The three-alignment lens (regulatory, operational, infrastructural) is offered as a design conjecture for regulated-classification tasks; in this study, operational alignment (vocabulary reach) was the binding constraint.

B. Limitations

Scope first: Annex I routing is not implemented (minimal outcomes are qualified accordingly), and of Article 50 only 50(1), 50(2) and the deep-fake limb of 50(4) are encoded; Art. 50(3) and the public-interest-text limb of 50(4) are not, so an Art. 50(3) duty is never surfaced. Intake is unstructured free text; the rule set is a first, reviewable encoding; reference labels are author-assigned and, so far, reproduced by practitioners rather than adjudicated by lawyers; the reference labels are predominantly single-tier, so Article 50 concurrency is demonstrated by construction rather than measured; the comparator is an internal floor; and off English the engine fails silently at 58-83%, so a translation front-end or per-language rule sets are a required layer for Union-wide use, not an optional extension. The perfect-residual figure of Section IV-C is an analytical ceiling and is evidence for none of this. These bound the claim precisely: TierGuard is a measured baseline and evaluation protocol for free-text AI-Act triage, not a legal classifier and not a deployable screen.

VI. CONCLUSION

Risk-tier determination is an entry-point control of the EU AI Act and still largely a manual judgement. We presented a lexical, deterministic triage-and-abstention baseline for Art. 5, Annex III and Art. 50 that separates workflow escalation from legal conclusion via a first-class indeterminate outcome, binds every candidate decision to its provision, and is evaluated with abstention-aware, error-direction metrics on post-freeze and regulator-derived sets, with its lexical limits measured adversarially and multilingually rather than asserted away. The result is a floor, not a screen: 40% of post-freeze duties pass it silently, its coverage stops at Annex I and at Art. 50(3), and its reference labels have been reproduced by practitioners but not yet adjudicated by legal experts. The next steps follow: the pre-specified expert adjudication round, a translation layer or per-language rules, the remaining Article 50 paragraphs, and published semantic and LLM-based classifiers run on the released benchmark under the same metrics. Baseline, benchmark suite, decision-record schema and study package are public for that purpose.

VII. ETHICS AND REPRODUCIBILITY

The evaluation uses synthetic and public-knowledge descriptions; no personal data are processed. Outputs are candidate or review-required records pending human legal confirmation, not legal advice, and no substitute for conformity

assessment. The engine is fully specified by its versioned rules and outcome model; deterministic figures follow from it and the fixed corpora. Rules, all benchmark sets and labels, the attack generator, harnesses, the SHA-256 manifest, an MIT licence, a provenance file giving per-source attribution and access dates, and one-command reproduction scripts ship as a package that is public at osf.io/q67sx (doi: [10.17605/OSF.IO/Q67SX](https://doi.org/10.17605/OSF.IO/Q67SX)). The two-annotator study (both guides, scorer, agreed labels, unresolved rows) ships as a separate package there, and a browser implementation of the same rule set, replayed against the artifact on every benchmark row, runs at tasnif.live/eu-classifier. The command reproduces the deterministic classification metrics. Latency is not one of them: it is remeasured on every run and written to `runtime_metrics.json`, which ships with the package but is deliberately outside the SHA-256 manifest, since a machine-dependent number cannot reproduce byte-for-byte. The file records the author’s environment alongside the figure so a different measurement elsewhere is explainable rather than a contradiction.

REFERENCES

- [1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024.
- [2] Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), OJ L, 2026/1744, 24.7.2026; in force 27 July 2026.
- [3] European Commission AI Office, "EU AI Act Compliance Checker (beta)," AI Act Service Desk, accessed July 2026.
- [4] T. Schuster, L. Waidehlich, A. Schneider, and M. Lambert, "Risk classification and compliance of AI systems under the EU AI Act," in Proc. AMCIS, 2025.
- [5] T. B. Momcilovic, D. Balta, B. Buesser, G. Zizzo, and M. Purcell, "Knowledge-augmented reasoning for EU AI Act compliance and adversarial robustness of LLMs," [arXiv: 2410.09078](https://arxiv.org/abs/2410.09078), 2024.
- [6] A. Davvetas, M. Papademas, X. Ziouvelou, and V. Karkaletsis, "AI Act evaluation benchmark: an open, transparent, and reproducible evaluation dataset for NLP and RAG systems," [arXiv: 2603.09435](https://arxiv.org/abs/2603.09435), 2026.
- [7] A. Buscemi, T. Deckenbrunnen, F. Kabir, K. Mishchenko, and N. Mowla, "Assessing high-risk AI systems under the EU AI Act: from legal requirements to technical verification," [arXiv: 2512.13907](https://arxiv.org/abs/2512.13907), 2025.
- [8] D. Golpayegani, H. J. Pandit, and D. Lewis, "AIRO: an ontology for representing AI risks based on the proposed EU AI Act and ISO risk management standards," in Towards a Knowledge-Aware AI (SEMANTICS), IOS Press, 2022, pp. 51-65. doi: [10.3233/ssw220008](https://doi.org/10.3233/ssw220008)
- [9] D. Golpayegani, H. J. Pandit, and D. Lewis, "To be high-risk, or not to be, semantic specifications and implications of the AI Act’s high-risk AI applications and harmonised standards," in Proc. ACM FAccT, 2023, pp. 905-915. doi: [10.1145/3593013.3594050](https://doi.org/10.1145/3593013.3594050)
- [10] D. Golpayegani, H. J. Pandit, D. O’Sullivan, and D. Lewis, "Semantic frameworks to support implementation of the EU AI Act," Computer Law & Security Review, vol. 61, art. 106315, 2026, doi: [10.1016/j.clsr.2026.106315](https://doi.org/10.1016/j.clsr.2026.106315).
- [11] R. C. Ugarte, M. Á. P. Guisado, A. B. de Jesús, and J. M. M. López, "Making AI compliance evidence machine-readable," [arXiv: 2604.13767](https://arxiv.org/abs/2604.13767), 2026.
- [12] M. Hanif, M. Constantino, and M.-T. Sekwenz, "Tough decisions? Supporting system classification according to the AI Act," in Frontiers in Artificial Intelligence and Applications (JURIX), 2023. doi: [10.3233/FAIA230987](https://doi.org/10.3233/FAIA230987)

- [13] A. Kott, A. Roessler, and R. Gross, "AI regulates AI: an artifact for automated risk assessment," in INFORMATIK 2025, Gesellschaft fuer Informatik, 2025. doi: [10.18420/inf2025_117](https://doi.org/10.18420/inf2025_117)
- [14] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75-105, 2004. doi: [10.2307/25148625](https://doi.org/10.2307/25148625)
- [15] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," J. Management Information Systems, vol. 24, no. 3, pp. 45-77, 2007. doi: [10.2753/mis0742-1222240302](https://doi.org/10.2753/mis0742-1222240302)
- [16] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: a framework for evaluation in design science research," European J. Information Systems, vol. 25, no. 1, pp. 77-89, 2016. doi: [10.1057/ejis.2014.36](https://doi.org/10.1057/ejis.2014.36)
- [17] European Commission, "Guidelines on prohibited artificial intelligence practices," C(2025) 5052 final, 2025 (adopted; interpretive, non-binding on courts).
- [18] European Commission, "Draft Commission guidelines on the classification of high-risk AI systems," 19 May 2026 (non-binding draft; targeted consultation closed 23 Jul. 2026; final guidelines pending, expected by end 2026).

APPENDIX A RULE INVENTORY

Table III is the rule inventory: what each rule decides and on what authority. The *Trigger* column paraphrases the provision; the match expressions themselves, the complete lexical scope of the rule set, are in the artifact’s `ruleset.py`, and an example decision record is emitted by `passport.py`. All rules but Hw are strong (a match yields a candidate outcome); Hw is the deliberate weak net that routes a regulated-context signal to INDETERMINATE rather than tiering it down. Art. 50(3), the public-interest-text limb of Art. 50(4), and Annex I routing are not represented.

TABLE III
RULE INVENTORY (V0.3.0, PINNED TO REGULATION (EU) 2024/1689 AS AMENDED BY REGULATION (EU) 2026/1744). TRIGGERS ARE PARAPHRASES OF THE PROVISION; THE MATCH EXPRESSIONS ARE IN THE ARTIFACT’S `RULESET.PY`. ONLY Hw IS WEAK: IT NEVER FINALISES A TIER, IT ROUTES TO INDETERMINATE.

ID	Provision	Trigger
P1	Art. 5(1)(a)	subliminal / manipulative techniques
P2	Art. 5(1)(b)	exploiting vulnerabilities
P3	Art. 5(1)(c)	social scoring
P4	Art. 5(1)(d)	predictive policing (profiling only)
P5	Art. 5(1)(e)	untargeted facial scraping
P6	Art. 5(1)(f)	emotion recognition at work / school
P7	Art. 5(1)(g)	sensitive biometric categorisation
P8	Art. 5(1)(h)	real-time remote biometric ID (law enforcement)
H1	Annex III(1)	biometrics (non-prohibited)
H2	Annex III(2)	critical infrastructure
H3	Annex III(3)	education and vocational training
H4	Annex III(4)	employment and worker management
H5	Annex III(5)	essential services (benefits, credit, insurance)
H6	Annex III(6)	law enforcement
H7	Annex III(7)	migration, asylum and border control
H8	Annex III(8)	justice and democratic processes
P9	Art. 5(1)(ba)	non-consensual intimate imagery [from 2 Dec 2026]
P10	Art. 5(1)(bb)	CSAM generation [from 2 Dec 2026]
Hw	Annex III (context)	WEAK: regulated context, use-case unresolved -> INDETERMINATE
T1	Art. 50(1)	interacts with natural persons (provider duty)
T2	Art. 50(2)	generative / synthetic content (provider duty)
T3	Art. 50(4)	deep-fake limb only (deployer duty)