
Capabilities, Not Just Domains: A Minimal Amendment for Agentic AI Risk in Brazil's PL 2338/2023¹

Leo Arruda
GEP Riscos Globais, Unifesp

With
Apart Research

Abstract

Brazil's AI bill (PL 2338/2023) classifies risk by application domains rather than system capabilities. To evaluate this approach, I checked the bill's 80 articles against an eight-dimension framework for agentic risk.

The analysis reveals significant gaps for currently deployed systems. Across six near-term dimensions, the bill's Coverage Score — a heuristic indicator, not a calibrated metric — is just 25% (3/12) after an adversarial robustness check, with zero dimensions covered unconditionally. The two frontier dimensions — autonomous replication and multi-instance coordination — score 0/4; I read this as a stress test of statutory adaptability, not evidence of imminent risk. A narrow, four-dimensional comparison with the EU AI Act highlights a shared oversight-gating pattern, but finds the EU framework explicitly names autonomy and tool use as systemic-risk criteria, whereas PL 2338 does not.

Testing four deployment scenarios confirms that near-term protections remain domain-contingent rather than capability-contingent. Because these gaps interact — interruptibility and traceability act as force multipliers for risk duration and discoverability — fixing the initial classification mechanism yields high leverage.

I propose a narrowly scoped amendment to Articles 15 and 12, deliberately leaving frontier capabilities outside binding text: a single capability-based criterion targeting autonomy, multi-step planning, and tool-mediated action lets the statute capture near-term agentic risk through a tractable legislative fix, with mandatory preliminary assessment for systems that cross the threshold.

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

Brazil's AI bill (PL 2338/2023) [1] was approved by the Senate in December 2024 and is under review in the Chamber of Deputies' Special Committee (rapporteur: Deputy Aguinaldo Ribeiro, PP-PB), positioning it as one of the first comprehensive AI governance frameworks in Latin America. The Committee has held 31 meetings and heard 180 witnesses but has not yet issued a parecer; the floor vote, originally expected by the end of 2025, has been pushed into 2026 [19]. The bill therefore remains genuinely open to amendment, which is the practical stake of this paper. Like the EU AI Act that inspired it, PL 2338 classifies AI systems into risk tiers based on application domain — an approach designed for single-purpose deployed systems making discrete predictions or recommendations.

That assumption is increasingly strained and outdated. Deployed AI agents now plan, use tools, and execute multi-step, consequential actions with limited human supervision [2]. Browser-based and enterprise agents are already the most agentic systems in active deployment [4], and their ability to autonomously execute tasks across web, software, and physical environments creates security and safety risks distinct from both traditional AI safety and conventional software security [7]. Already in 2024, a consensus paper by Bengio, Hinton, and colleagues in Science concluded that present governance initiatives barely address autonomous systems [3]. Here I ask whether, and to what extent, this gap holds specifically for PL 2338/2023, and what it would mean in practice for AI deployers and regulators in Brazil.

Three structural assumptions, not isolated drafting gaps. The mismatch this paper documents traces back to three implicit assumptions a domain-based, risk-tiered framework makes about what AI risk is: that risk tracks where a system is deployed rather than what it can do; that risk is determinable once, before market introduction, rather than continuously as a system is fine-tuned or scaled; and that risk is a property of individual systems in isolation rather than something that can emerge from how systems interact. All three held reasonably well for the single-purpose, predictive systems the framework was built around, and break down for agentic ones. The first is the one most directly testable against statutory text, and my eight-dimension coverage framework operationalizes it empirically; the second and third are treated as structural limitations on this paper's proposed fix, not only on the status quo (§5).

The core argument. PL 2338/2023 classifies AI systems by what domain they operate in — a closed, twelve-item list (Art. 14) — not by what they can do. Agentic systems already in commercial deployment plan, use tools, and take consequential action regardless of which domain they sit in, so a domain-based test systematically misses them. A system can have full autonomy, no human override, and no audit trail, and still fall outside every one of the bill's strongest protections simply because its deployer's sector is not on the list — the bill's default treatment of agentic AI deployed today (§4.1–§4.3), not a hypothetical edge case.

The fix does not require rebuilding the statute: Art. 15 already gives the SIA — the Sistema Nacional de Regulação e Governança de Inteligência Artificial — authority to add classification criteria by regulation. A single capability-based criterion there, scoped to autonomy, multi-step planning, and tool/API-mediated consequential action, closes most of the near-term gap at once, since interruptibility and traceability attach automatically once a system is classified high-risk (§ 4.4.1, § 4.5–§ 4.6).

My main contributions are:

1. I develop an eight-dimensional coverage score for agentic AI risk, with an explicit coding rubric grounded in the agentic-AI-safety literature, and check the full 80-article text of the Senate-approved substitutivo against it.
2. I test the bill's risk-classification logic (Art. 14) against four realistic deployment scenarios, showing coverage is domain-contingent rather than capability-contingent.
3. I provide a limited, four-dimensional comparison with the EU AI Act's systemic-risk regime (Arts. 51–56, Annex XIII), suggesting the gap is structural to risk-tiered regulation generally, not a Brazil-specific drafting accident.
4. As a secondary, stress-test finding — not a claim of imminent risk — I identify two provisions with no PL 2338 analogue: autonomous replication/self-modification, and aggregated multi-instance risk. I also note a national-defense carve-out (Art. 1º, § 1º, II) excluding defense-relevant autonomous agents from the law's scope entirely.
5. I propose a narrowly scoped amendment — a capability-based criterion limited to autonomy, multi-step planning, and consequential action, added to Art. 15 rather than reopening the Art. 14 domain list — and assess its tractability against the Senate's amendment history (Parecer nº 207/2024-PLEN/SF) and its leverage via a cross-dimension impact analysis (§ 4.4.1).
6. I argue the gap follows from the three structural assumptions named above, and that the coverage framework tests only the first of the three, leaving the second and third as limitations on this paper's own proposed fix, not only on the status quo (§ 5).

2. Related Work

PL 2338/2023 and the existing critique. Civil society and legal analysts have engaged extensively with the bill, but along different axes than mine. Data Privacy Brasil produced a technical analysis ahead of the Senate floor vote [8], focused on rights and procedural safeguards; the Coalizão Direitos na Rede has argued industry lobbying weakened governance obligations relative to an earlier draft [9] — a capture-and-process critique, not a capability-coverage one. Other commentators argue the bill is stricter than the EU AI Act and disproportionately burdens smaller firms — a market-structure critique. None of this work asks whether the bill's risk

taxonomy is structurally suited to agentic AI specifically; my work addresses that question directly.

Agentic AI risk taxonomies. I ground my coverage dimensions in the agentic-AI-safety literature rather than constructing them ad hoc. Chan et al. (2023) identify underspecification, directness of impact, goal-directedness, and long-term planning as key characteristics of agentic risk [2], motivating my planning, consequential-action, and autonomy-scaled-obligations dimensions; governance work treating safe interruptibility as a distinct risk pillar [6] anchors my interruptibility dimension. For autonomous replication, I draw on METR's original autonomous-replication-and-adaptation framing [10] and on RepliBench's subtask decomposition [11], whose authors find no credible near-term self-replication threat, though capability is improving rapidly. For multi-instance coordination, I draw on security taxonomies treating multi-agent interaction as a distinct attack surface [7][12]; recent agent-risk-assessment work [5] similarly notes the field lacks a unified, validated approach to operationalizing agent-level risk — the gap that makes a textual coverage audit, rather than assuming existing frameworks transfer cleanly to legal text, methodologically necessary here.

Regulatory reference point. The EU AI Act is relevant as PL 2338's acknowledged template: it explicitly names "level of autonomy and scalability" and "tool use" among the Annex XIII criteria for designating general-purpose models as posing systemic risk [13][14] — a more explicit capability-based trigger than anything in PL 2338 (§4). I do not attempt a full comparative analysis, which is outside what a single-author, two-day study can responsibly support; I use this single point of contrast only to test whether my central finding is Brazil-specific or structural (§4.2).

3. Methods

Primary source. I use the Senate-approved *substitutivo* of PL 2338/2023 — the *Redação para o turno suplementar*, Parecer nº 208, de 2024-PLEN/SF, approved 10 December 2024 — retrieved from the Senado Federal legislative repository [1]. This 80-article consolidated text supersedes the original 45-article bill filed by Senator Rodrigo Pacheco in 2023; readers verifying citations below should confirm they are viewing the Parecer 208 text, which has materially different article numbering.

The Chamber of Deputies' Special Committee held 12 of its 31 meetings as dedicated public hearings between May and September 2025 and may still amend this text; the rapporteur's *parecer* has not yet been issued. My analysis is anchored to the Senate-approved version — the most recent stable, citable text — and findings should be revisited against the final enacted text once available.

Step 1 — Framework construction. I defined eight coverage dimensions:

Table 1 — The eight coverage dimensions

#	Dimension	Question coded
D1	Autonomy-scaled obligations	Do governance obligations scale with a system's degree of autonomy, ex ante?
D2	Tool/API-mediated action	Does the bill govern risk arising from a system's use of external tools/APIs to take action?
D3	Extended/multi-step planning	Does the bill address risk from a system pursuing goals across multiple steps without per-step human sign-off?
D4	Consequential real-world action	Do obligations scale with whether a system executes consequential actions, not just produces outputs/recommendations?
D5	Interruptibility / human override	Can a human stop or override the system's operation, as a matter of right?
D6	Action traceability / audit logging	Must the system's autonomous actions be logged and attributable?
D7	Autonomous replication / self-modification	Does the bill address a system's capacity to replicate itself or modify its own behavior post-deployment?
D8	Distributed / multi-instance coordination (same system)	Is there a mechanism to assess aggregate risk from multiple coordinated instances of the same system?

D1–D6 are motivated by Chan et al.'s (2023) characteristics of agentic risk [2] and by interruptibility/oversight framing in the agentic-AI governance literature [6]. D7 draws on the autonomous-replication-and-adaptation threat model from frontier-lab evaluations [10][11]; RepliBench's evaluators find no credible near-term replication threat [11], so D7 functions here as a stress test of statutory adaptability, not a claim that replication is imminent in Brazil — a more equivocal subsequent assessment narrows that gap somewhat (§4.4(d), [16]). D8 draws on security taxonomies treating multi-agent interaction as a structurally distinct attack surface [7][12]: a set of instances, each individually below a risk threshold, can produce aggregate harm no per-instance assessment would catch, breaking the assumption throughout Arts. 12–18 that "a system" is the correct unit of analysis. Unlike D7, I treat D8 as near-term rather than frontier: coordinated swarms of unremarkable agents (fraud networks, disinformation campaigns) are

observed today, which is why §4.4.1 treats it separately from D7 despite their shared tier in Table 2.

The eight dimensions are deliberately conservative, not exhaustive — selected for recurring across independent agentic-risk taxonomies and mapping cleanly onto a yes/no statutory-coverage question, rather than enumerating every capability the literature associates with agentic risk. Excluded candidates (deception, persuasion, situational awareness, offensive cyber capability) and the rationale for equal weighting across dimensions are detailed in Appendix C.1.

Step 2 — Coding rubric. Each dimension was scored against the bill's text using the following explicit criteria:

- **Unconditional (2):** the bill contains a provision that directly and generally addresses this dimension, applying regardless of risk classification, sector, or other gating condition.
- **Conditional (1):** the bill addresses this dimension, but only contingent on a separate gating condition being met first (e.g., the system must already be classified "high risk" via the closed domain list in Art. 14, or the obligation applies only to public-sector deployers).
- **Absent (0):** no provision, explicit or conditional, addresses this dimension under the current text.

I do not code principle-level text — the "fundamentos" and "principios" of Arts. 2^o–3^o — as coverage. Principles inform interpretation but do not themselves create an enforceable duty under Art. 50, and give an affected person no mechanism to invoke them. This holds regardless of how directly a principle's wording touches the subject matter; **only operative provisions are coded**. Appendix A flags each case where a principle-level candidate exists and was excluded on this basis.

I conducted a full manual close-reading of all 80 articles, recording for each dimension the supporting article(s), operative language, and resulting score. The four most contestable codings were additionally stress-tested with an adversarial re-read (§4.1). The full article-by-article mapping is in Appendix A.

Coverage Score. I compute a blended Coverage Score = $(\Sigma \text{ dimension scores}) / 16$, where 16 is the maximum possible (8 dimensions $\times$ 2). This should be read as a heuristic summary indicator, not a validated measurement scale: both the 2/1/0 weighting and the equal weighting across dimensions are modeling choices made for replicability and legibility, not principled calibrated values. Because an unweighted blend treats all eight dimensions as equally consequential regardless of real-world prevalence, I also report a **near-term Coverage Score** (D1–D6, max 12) covering capabilities already observed in deployed systems [4], and a **frontier Coverage Score** (D7–D8, max 4) covering capabilities not yet credibly demonstrated in evaluated models [11]. The tiered scores, not the blended one, are the primary basis for the recommendation in §4.5.

Step 3 — Case studies. I constructed four deployment scenarios reflecting categories of agentic systems already observed in production — personal/consumer, developer/coding,

enterprise/customer-facing, and multi-instance — informed by the 2025 AI Agent Index [4]. Each was traced through Art. 14 (high-risk classification) and Art. 4°, III / Arts. 29–33 (general-purpose AI pathway) to determine which obligations would be triggered.

Step 4 — Limited external benchmark. To test whether my findings are Brazil-specific or structural, I coded the same eight dimensions against the EU AI Act's systemic-risk chapter (Arts. 51–56, Annex XIII) and human-oversight provision (Art. 14), using public primary and secondary sources [13][14][15]. This benchmark is intentionally narrow — I did not code the EU Act's full 113 articles, and only four of eight dimensions were checked — and should be read as directional, not comprehensive.

Limitations of method (detailed further in Section 5): single coder, with no independent inter-rater check — an adversarial self-administered re-read on the four most contestable codings (§4.1) partially mitigates for this; legal text read by a non-lawyer, without formal validation by a Brazilian digital-law specialist within the two-day hackathon timeframe; the EU AI Act benchmark in Step 4 is narrower in scope than the primary PL 2338 audit.

4. Results

4.1 Coverage Matrix

Table 2 shows the scored coverage of PL 2338/2023 across all eight dimensions, following the rubric in Section 3 and incorporating the adversarial robustness check described below.

Table 2 — Per-dimension coverage

Dimension	Score	Tier
D1. Autonomy-scaled obligations	Absent (0)	A — near-term
D2. Tool/API-mediated action	Absent (0)	A — near-term
D3. Extended/multi-step planning	Absent (0)	A — near-term
D4. Consequential real-world action	Conditional (1)	A — near-term
D5. Interruptibility / human override	Conditional (1)	A — near-term
D6. Action traceability / audit logging	Conditional (1)	A — near-term
D7. Autonomous replication / self-modification	Absent (0)	B — frontier

D8. Distributed / multi-instance coordination Absent (0) B — frontier
(same system)

All positive coverage (D4, D5, D6) is Conditional — gated behind Art. 14 risk classification or sector. D1, D2, D3, D7, and D8 are Absent. No dimension reaches Unconditional (2). Full article-by-article justification is in Appendix A.

The bill's blended Coverage Score is 3/16 (19%); split by tier, near-term dimensions (D1–D6) score 3/12 (25%) and frontier dimensions (D7–D8) score 0/4 (0%). No dimension in either tier is covered without a gating condition attached. The tiered view shows near-term coverage is meaningfully higher than the blended indicator implies — though still entirely conditional — while frontier coverage is total absence, not partial. This distinction matters for sequencing remediation: §4.5 proposes a fix targeted at the structural cause of the near-term gap (the absence of any capability-based classification criterion), while treating frontier dimensions as a secondary adaptability question rather than an equally urgent legislative target today.

Because this is single-coder analysis, I stress-tested the four most contestable scores (D1, D2, D4, and Case Study 3, §4.3) with a deliberate adversarial re-read — actively arguing against my own initial coding rather than confirming it. This produced one revision: D1 (autonomy-scaled obligations) moved from Conditional to Absent, since Art. 36's liability factor answers an ex-post question rather than the ex-ante one the dimension asks (Appendix A). D2 and the Case Study 3 finding held; D4 held but is the more borderline of the two. The check moves the score down, not up — the more trustworthy direction for a self-administered test — and the qualitative finding holds across the resulting sensitivity range (17–33% near-term coverage, depending on how D1 and D4 are treated): no near-term dimension is unconditionally covered, and the amendment proposed in §4.5 is unaffected by which exact score within that range a reader accepts, since it targets the same cluster of dimensions regardless of how each is coded. This is not a substitute for independent second-coder validation (§5).

4.2 External Benchmark (EU AI Act, limited scope)

Table 3 — External benchmark: PL 2338/2023 vs. EU AI Act (four dimensions)

Dimension	PL 2338/2023	EU AI Act
D5. Interruptibility	Conditional (1) — Art. 8°, gated to Art. 14 "high risk"	Conditional (1) — Art. 14, gated to Annex III high-risk classification [15]

D2. Tool/API use	Absent (0) — Art. 32's cooperation duty governs API as a distribution channel, not tool-use-driven action	Conditional (1) — "level of autonomy and tool use" is a named Annex XIII factor for systemic-risk designation, but only for models meeting that threshold [13][14]
D7. Replication/self-modification	Absent (0)	Conditional (1) — "level of autonomy and scalability" is a named Annex XIII factor [13][14], though no provision targets autonomous replication specifically
D8. Distributed/multi-stance (same system)	Absent (0)	Not identified in the sources reviewed for this limited benchmark — treated as tentatively Absent (0); a full Annex III/Annex XIII reading would be needed to confirm

This narrow, four-dimension check finds the same oversight-gating pattern in both regimes — the EU AI Act's strongest human-oversight right (Art. 14) is, like PL 2338's Art. 8º, gated behind a closed domain list (Annex III), not a capability test. It also finds that on autonomy and tool use specifically, the EU AI Act's Annex XIII does name these as explicit systemic-risk designation criteria for general-purpose models [13][14], where PL 2338 currently has no equivalent. PL 2338 does have its own systemic-risk concept (Art. 4º, XXX), gating the heavier SIAPG (sistema de IA de propósito geral, general-purpose AI system) obligations in Art. 30, paralleling the EU AI Act's — but unlike Annex XIII, it names no contributing factors at all, not autonomy, not tool use, not compute, leaving the threshold entirely to future regulamentação. This sharpens the comparison rather than changing it: both regimes have a systemic-risk concept, but only the EU's currently names capability-based criteria.

Read together, at the confidence level the evidence actually supports: domain-gating on human-oversight rights is not unique to Brazil's drafting (one data point, but a real one), and naming capability as a regulatory trigger is achievable inside a risk-tiered framework — the EU has already done it for two of my four checked dimensions — rather than a purely aspirational ask. I do not extend this to a general claim about risk-tiered regulation beyond these two systems; that would require coding the EU Act's remaining four dimensions and its full Annex III/XIII text directly (Section 5, Future Work).

4.3 Case Study Outcomes

Table 4 — Case study classification outcomes

Case study	Falls within Art. 14 domain list?	Triggers high risk status?	Triggers SIAPG duties (Arts. 29–33)?	Triggers Art. 8° human-supervision requirement?
1. Personal purchasing agent (compares prices, executes purchase autonomously)	No	No	Yes, if built on a general-purpose model	No
2. Autonomous coding agent (writes, tests, deploys code to production)	Only if the operator runs critical infrastructure (Art. 14, I)	Conditional	Yes	Conditional
3. Customer-service agent with refund authority	No (currently) — contingent on future "essential services" sectoral regulation under Art. 14, IV / Art. 15	No (currently)	Yes	No (currently)
4. Multi-instance deployment (N parallel copies of the same agent, each performing the same task)	No — assessment is per-system	No (no aggregation mechanism)	Yes, per instance	No

Case 2's mapping to Art. 14, I (critical infrastructure), is conditioned on "risks to physical integrity" and the system being a determining factor for the outcome. Case 3 maps to Art. 14, IV (essential services) but is not currently triggered, since no sectoral regulation yet designates commercial refund or eligibility decisions "essential" — a present-tense "No," contingent on future Art. 15 sectoral rulemaking.

4.4 Key Findings

- (a) No dimension is unconditionally covered, and near-term coverage (25%) is meaningfully better than frontier coverage (0%). All positive coverage in PL 2338 is

Conditional, gated either behind Art. 14's closed domain list or behind a public-sector-only application (Art. 23, I).

- **(b) The SIAPG pathway (Arts. 29–33) is the bill's only consistent catch-all for agentic systems, but it regulates documentation about the model, not the agent's run-time behavior.**
- **(c) The bill has no mechanism for aggregate, multi-instance risk (D8), and the EU AI Act, on the evidence reviewed, may share this gap.** Unlike D7, this gap is not capability-gated (§3) — coordination, not capability, is the binding constraint — making D8 the more concretely exploitable of the two frontier-tier dimensions today despite sharing that tier's label.
- **(d) No provision addresses autonomous replication or self-modification post-deployment (D7)** — included here as a stress test of statutory adaptability rather than a claim of current risk in Brazil specifically; the EU AI Act, by contrast, names autonomy as a systemic-risk factor, even if not specifically a replication provision.

This framing should be read against fast-moving evidence: METR's Frontier Risk Report (May 2026), based on direct, non-public access to the strongest internally-deployed models at four frontier labs (Anthropic, Google, Meta, and OpenAI), found these models plausibly already had the means, motive, and opportunity to start small, non-robust rogue deployments as of March 2026 [16] — a meaningfully closer call than the "no credible near-term threat" baseline RepliBench originally scored this dimension against [11]. This does not show autonomous replication is an imminent risk in Brazil specifically — it concerns internal models at non-Brazilian labs, not commercial products — but D7's absence now reads as a live adaptability question rather than a purely hypothetical one. The bill's text has not changed, so D7 stays Absent (0).

4.4.1 Plausible Impact and Cross-Dimension Interactions

The Coverage Score (§4.1) is a coverage indicator, not a harm estimate: it does not say what follows, in practice, from each gap remaining open. The tiered scoring (§3) corrects for prevalence — how common the underlying capability is in deployed systems — but not for severity, and the case studies (§4.3) report classification outcomes without stating their consequences. Table 5 makes that layer explicit, anchored to the case studies and literature already cited rather than to new scenarios, and should be read throughout as the author's structured judgment rather than a measured finding.

Table 5 — Plausible impact by dimension

Dim	Harm channel	Severity	Interaction
-----	--------------	----------	-------------

D1	No preventive incentive — autonomy is weighed only after harm occurs, in litigation (Art. 36)	Moderate — weakens deterrence; does not itself create harm	Depends on D6: without a log, there is nothing for a judge to weigh
D2	Case 1 — autonomous purchasing/transaction agents carry no obligation on the action-taking step itself	Moderate, scaling with deployment volume: small per-incident errors become aggregate harm at scale	Compounds with D5: without interruptibility, a flawed batch of actions cannot be halted mid-execution
D3	Zero coverage; cumulative goal drift across steps, each locally defensible, with no step that triggers review	Potentially high — the only dimension where nothing in the statute would prompt scrutiny of trajectory rather than endpoint	Compounds with D6: no record of intermediate reasoning makes post-hoc D1 liability analysis difficult to construct
D4	Case 3 — consequential decisions (financial, access to services) outside the domain list face no obligation scaled to consequence	Moderate to high, depending on the action class	Same as D2
D5	Not a standalone harm — a <i>duration multiplier</i> on how long any other gap's harm runs before a human supervisor is required to be positioned to intervene	Multiplier, not independent severity	Touches nearly every other dimension
D6	Public-sector-only logging (Art. 23, I) leaves private-sector harm undiscoverable and unattributable	Low as a standalone harm, high as a force-multiplier on the unaccountability of every other gap	Foundational to D1 and D3
D7	See § 4.4(d); capability-dependent — currently "small, non-robust" per third-party assessment of frontier labs' internal models, trending toward robust [16]	Low-moderate today, time-sensitive upward	—

D8	Not capability-dependent (§3): coordinating many ordinary, already-available agentic instances is the operating model behind automated fraud networks and astroturfed political amplification — orchestration, not a technical advance	Already realized in adjacent domains today, not contingent on any future capability threshold	Compounds with D6: a coordinated swarm operating below any single instance's risk threshold leaves no aggregate trace for any one deployer's log to capture
----	---	---	--

Two findings follow that the per-dimension scoring alone does not surface. First, D5 and D6 are not peers of D1–D4 — they are multipliers: a system that acts consequentially without being interruptible (D5) has its harm extended in time, and one that acts without being logged (D6) has that harm rendered undiscoverable. This also disables D1's only enforcement mechanism, since ex-post liability under Art. 36 requires a record to litigate against. The near-term gap is thus better described as four action-and-planning gaps (D1–D4) whose practical consequences are determined by two structural gaps (D5, D6) the bill happens to gate on the same classification test. Second, this is why the Art. 15, XI amendment (§4.5) is worth more than its narrow scope suggests: because D5 attaches automatically once a system reaches high-risk status, a single capability-based trigger does not just close D1–D4, it activates D5 as a side effect for every system it captures. D6 needs a separate fix, which §4.5's companion provision supplies.

Who bears the cost. Mapped onto the case studies (§4.3), these gaps describe specific, foreseeable users, not abstract categories: a consumer whose purchasing agent (Case 1) executes an erroneous transaction has no PL 2338-mandated supervisor positioned to interrupt it (D5) and no obligation on the transaction itself (D2) — though payment-infrastructure safeguards likely cap worst-case severity independent of this statute. A person wrongly denied by an automated refund system (Case 3) has no PL 2338 log to point to (D6) and no forum weighing the system's autonomy beforehand — D1 is Absent, not merely delayed (Appendix A).

One separately-administered avenue exists outside PL 2338: LGPD Art. 20 gives data subjects a right to review of decisions taken "unicamente" by automated processing [17], though commentary documents companies narrowing this trigger via nominal human confirmation [18] — the same pro-forma-approval pattern §4.5's amendment is built to foreclose.

The asymmetry is structural, not incidental: coverage depends on which of Art. 14's twelve domains a system falls into, not on what it can do to the person interacting with it, so two functionally identical systems can receive entirely different protection based on their deployer's sector label. Closing D1–D6 via Art. 15, XI restores the missing second criterion — capability — without displacing the first.

4.5 Proposed Amendment

The gap identified above is structural but narrow, and — per the tier split in §4.1 — most consequential for capabilities already deployed rather than frontier ones. PL 2338's risk-classification mechanism (Arts. 12–16) has no criterion tied to a system's own capability profile, only to the affected person's circumstances.

Reopening the closed domain list in Art. 14 would relitigate a fight already settled across multiple *emenda* cycles in the Senate — Parecer nº 207/2024-PLEN/SF records the rejection of capability-adjacent amendments (e.g., Emenda nº 217, reinserting internet-application systems into Art. 14) as "excessively generic." That precedent is itself a reason to keep this proposal narrow: a capability criterion broad enough to read as a frontier-safety wish list risks the same fate. One might object that criterion XI's own threshold language is exactly the kind of open-textured standard that doomed Emenda 217 — but the rejection turned on *unbounded scope* (a sector re-admission with no behavioral test attached), not on vocabulary. Criterion XI instead names three conjunctive, independently observable behavioral properties that must co-occur, with an explicit pro-forma-approval floor closing the obvious self-exemption route (Appendix C.2). The residual risk sits one level downstream, at calibration rather than drafting — how loosely the SIA reads "multi-step planning" or "estado da arte" once it exercises rulemaking authority — a problem this proposal mitigates with a consultation-step recommendation, not one it eliminates outright.

A tractable insertion point is Art. 15, which already grants the SIA rulemaking authority to update the high-risk list using ten listed criteria (I–X) — all currently impact-based, none capability-based.

Proposed new subitem ("inciso"), Art. 15 — scoped to the three dimensions with the clearest empirical grounding in currently deployed systems [4]: autonomy (D1), multi-step planning (D3), and tool/API-mediated consequential action (D2, D4):

XI – o grau de autonomia do sistema de IA, incluindo sua capacidade de planejamento multietapa sem aprovação humana substantiva por etapa — desconsiderando-se, para esse fim, a aprovação meramente pro forma ou que não importe em efetiva possibilidade de revisão, modificação ou veto da ação pelo agente humano — e de execução de ações consequenciais mediante uso de ferramentas ou interfaces de programação de aplicação (API), conforme avaliação técnica baseada no estado da arte.

(English gloss, for reference only — the operative text is the Portuguese above: "the degree of autonomy of the AI system, including its capacity for multi-step planning without substantive human approval at each step — disregarding, for this purpose, approval that is merely pro forma or that does not entail a genuine possibility of review, modification, or veto of the action by the human agent — and for executing consequential actions via tools or application programming interfaces (APIs), per technical assessment based on the state of the art.")

This formulation is deliberately drafted against the pro-forma-approval failure mode: a deployer cannot satisfy the "human approval" lower bound — and thereby avoid the mandatory Art. 12 assessment below — merely by inserting a rubber-stamp checkpoint into an otherwise autonomous pipeline. This does not close the separate Art. 14, §ún. "intermediate technology" carve-out, a textually distinct provision, but the same "substantive human approval" anchor is available to inform its interpretation by analogy.

Why this combination, specifically. A general-purpose model (SIAPG) is not inherently risky; API/tool integrability alone is a property of most software. What changes the picture is the combination — broad capability, tool/API access, and multi-step planning without substantive human approval — which is what makes a system's actual behavior hard for an affected person or regulator to anticipate or verify after the fact, the same information asymmetry §4.4.1 documents. Requiring an upfront capability evaluation for systems matching this profile is a proportionate response to that asymmetry, not a presumption that any given system is unsafe — a recognition that, absent such a check, no one but the deployer is positioned to know whether it is. Two existing evaluation methodologies — METR's task-horizon approach and RepliBench's subtask decomposition — already give the SIA a replicable template to operationalize this rather than building one from scratch (Appendix C.3).

Proportionality for ordinary commercial systems. Criterion XI requires autonomy, multi-step planning, *and* consequential tool/API action together, so a conversational system that reasons across turns but takes no consequential action falls outside it regardless of length. The companion provision below caps the consequence of crossing the threshold at the bill's lightest tier — a mandatory preliminary assessment (Art. 12, here made mandatory rather than voluntary) plus action logging — not the full high-risk package, closer to LGPD Art. 38's impact-assessment pattern than to a novel compliance regime. That said, this depends on the SIA calibrating the threshold sensibly; a threshold set too loosely would fall disproportionately on smaller developers, the standard failure mode of any capability-based trigger (the EU AI Act's Annex XIII threshold faces the same problem, §4.2). I do not have a textual basis for asserting Art. 15 rulemaking already requires public consultation before a new criterion takes effect; if it does not, building in a consultation step — mirroring ANPD's practice for comparable LGPD rulemaking — would mitigate this risk without reopening the amendment's text.

I deliberately exclude D7 (autonomous replication) from this binding clause, even though the same underlying logic applies there, arguably more acutely (§4.4(d)): Art. 15's existing authority already lets the SIA extend criterion XI's "state-of-the-art" language to additional capability classes, including replication, without further statutory amendment — offered here as guidance for future SIA regulation, not text for the Special Committee to enact now. D8 is excluded for a different, more fundamental reason: it breaks the assumption that "a system" is the correct unit of regulatory analysis, so a per-system classification trigger cannot, by construction, reach the aggregate behavior of coordinated instances — closing it requires a structurally different mechanism, not a better-calibrated version of this one.

Companion provision, new § 7^a on Art. 12 (or equivalent insertion point): direct the SIA to treat a documented capability threshold under proposed Art. 15, XI as a trigger, making the Art. 12 preliminary assessment **mandatory** rather than voluntary, and extending the Art. 23-I logging obligation to any system meeting that threshold regardless of sector. This solves two problems: first, nothing currently compels self-assessment for a system outside the domain list, so a capability criterion alone would have nothing to attach to for systems that decline to self-classify; second, Art. 23-I's logging duty carries its own public-sector gate, so without this extension a newly-captured private-sector system would gain a human-supervision duty (D5) without the log (D6) that makes it accountable. This pairing targets the structural cause of the near-term gap without requiring a new high-risk category or reopening the domain list — a lower-friction ask than the alternatives already debated and rejected during the 2024 Senate process.

Table 6 — Anticipated Objections

Objection	Response	Elaborated at
Capability isn't risk. A highly autonomous system may be harmless.	Correct — criterion XI does not presume danger. It targets an information asymmetry: once autonomy, tool/API access, and multi-step planning co-occur, no one outside the deployer can verify after the fact whether the system behaved safely. The criterion requires assessment, not prohibition.	Above, "Why this combination, specifically"
A capability trigger burdens innovation, especially smaller developers.	Real risk, addressed by design: the criterion is conjunctive (all three properties, not any one), caps consequence at the bill's lightest tier (mandatory preliminary assessment, not full high-risk status), and adds no new compliance category. The residual risk is SIA calibration, not statutory design — mitigated, not eliminated, by the consultation-step recommendation.	Above, "Proportionality for ordinary commercial systems"

The SIA could already read capability into Art. 15 without this amendment.	Art. 15's existing ten criteria (I–X) are uniformly impact-based; capability is not currently a recognized criterion type at all. A deployer disputing classification has a textual argument that capability sits outside the SIA's intended scope, and the SIA has no enumerated hook to answer it. Criterion XI's function is to make capability consideration textually authorized rather than contestable — which matters precisely in the disputed cases where an explicit criterion is needed most.	§ 4.6, "first point at which capability becomes a recognized criterion type"
---	---	--

4.6 Leverage, Timing, and Consequences

Leverage and timing. Two things make criterion XI worth more than its narrow scope suggests. First, per the interaction analysis in § 4.4.1, a single classification criterion, paired with the companion provision above, moves all six near-term dimensions (D1–D4 directly; D5 and D6 as triggered consequences) from gapped to operative for any system it captures, rather than fixing one dimension at a time. Second, architecturally: every other agentic-risk provision in this paper is downstream of a classification test that today takes only domain and sector as inputs (§ 4.1). Art. 15 is the one place the SIA already holds open-ended authority to add classification criteria by regulation. Criterion XI is the first point at which *capability* becomes a recognized criterion type alongside the existing impact-based criteria I–X. Once that type exists, extending it to further capability classes becomes a regulatory act under Art. 15's existing authority rather than a new amendment cycle.

The window to act is also more time-bound than the bill's two-year history suggests: the Senate's *emenda* cycle shows that once a position is rejected and recorded, as Emenda n° 217 was (§ 4.5), it becomes harder to reintroduce without relitigating the fight. The Chamber's Special Committee — 31 meetings in without a *parecer* — represents an unusually long-open window, not a stable baseline. Once the *parecer* is issued and the text moves toward a floor vote, the cost of introducing a new criterion rises substantially, for the same reason Art. 14's domain list has proven difficult to reopen.

Consequences of adoption versus the status quo. Table 7 states plainly what changes, dimension by dimension.

Table 7 — Consequences of adoption versus the status quo

Dimension	If not adopted (status quo persists)	If adopted (criterion XI + companion provision)
D1. Autonomy	Autonomy enters the law only after harm, as one factor in litigation (Art. 36) — no incentive operates before harm occurs	A documented capability threshold triggers mandatory pre-deployment assessment — autonomy is weighed before harm, not only in court
D2. Tool/API action	No obligation attaches to the action-taking step itself; errors on autonomous transactions scale uncaught with deployment volume	Systems combining tool/API access with autonomy must be assessed and logged before that capability scales into routine use
D3. Multi-step planning	Zero statutory foothold — the only dimension where nothing in the bill would prompt scrutiny of a system's trajectory, not just its endpoint	Closed directly: multi-step planning without substantive human approval is the core trigger condition, not a side effect
D4. Consequential action	Coverage is an accident of which Art. 14 domain a deployer happens to sit in, not of what the system does to the person it acts on	Consequential action via tools is captured on the basis of behavior, restoring capability as a second, independent classification axis
D5. Interruptibility (multiplier)	No statutory requirement that a human be positioned to intervene for any system newly exhibiting these capabilities; harm duration is unbounded by statute	A deployer-side human-supervision requirement activates automatically via the existing Art. 8º gate once high-risk status is triggered — no separate drafting required
D6. Traceability (multiplier)	Private-sector harm is undiscoverable and unattributable; the Art. 36 liability mechanism behind D1 has no record to litigate against	Companion provision extends Art. 23-I logging beyond the public sector — restores the evidentiary basis D1 and D5 both depend on

Put plainly: not adopting criterion XI does not mean the underlying risk goes unrealized — it means the bill has no statutory hook for it regardless of severity. Adopting it does not solve agentic AI risk in Brazil; it gives the statute, for the first time, a criterion that responds to capability at all, at the lightest compliance tier the bill provides (§4.5), and a mechanism through which further capability classes can be added by regulation rather than by relitigating Art. 14.

Quantified effect on the Coverage Score. Applying §3's rubric to the post-amendment text directly: near-term coverage (D1–D6) moves from 3/12 (25%) to 6/12 (50%); the blended score

moves from 3/16 (19%) to 6/16 (37.5%). Frontier coverage (D7–D8) is unchanged at 0/4, by design (§ 4.5). Table 8 shows the per-dimension movement.

Table 8 — Per-dimension score, before and after adoption

Dimension	Pre-amendm ent	Post-amendm ent	Why
D1. Autonomy	0 (Absent)	1 (Conditional)	A real ex-ante mechanism exists for the first time, gated to the conjunctive capability threshold
D2. Tool/API action	0 (Absent)	1 (Conditional)	Same threshold now reaches the <i>aplicador</i> 's own runtime tool/API use, which Arts. 29/32 do not
D3. Multi-step planning	0 (Absent)	1 (Conditional)	Criterion XI's core trigger condition; still gated, conjunctively
D4. Consequential action	1 (Conditional)	1 (Conditional)	A second, independent gated pathway is added alongside Art. 14's domain gate, but the categorical score does not move, since the dimension was already Conditional via the existing path
D5. Interruptibility	1 (Conditional)	1 (Conditional)	Art. 8 ^o still fires only on high-risk classification; criterion XI adds a second route into the same gate
D6. Traceability	1 (Conditional)	1 (Conditional)	The companion provision removes the public-sector restriction, but logging

remains gated — now to the capability threshold rather than to sector

D7. Replication	0 (Absent)	0 (Absent)	Deliberately excluded from binding text (§ 4.5)
D8. Multi-instance coordination (same system)	0 (Absent)	0 (Absent)	Excluded; a per-system trigger cannot reach aggregate, multi-instance behavior by construction (§ 4.5)

Two caveats apply, and both cut against overstating what the amendment does. First, every dimension that improves still lands at Conditional, not Unconditional — criterion XI does not remove gating from the bill's classification logic; it swaps the gate from domain-based to capability-based. The paper's finding that no dimension is covered unconditionally holds before and after adoption. Second, the categorical score understates D4–D6's improvement rather than overstating it: each gains a second, independent, or broader-population gate without that change registering as a higher number, since the rubric records whether a dimension is reached at all, not how many pathways reach it or how large the population behind the gate is.

Beyond the four dimensions it directly addresses, criterion XI's practical value is in moving statutory protection away from an accident of sector classification and toward what a system actually does to the person interacting with it — and, as one of the first comprehensive AI statutes in Latin America (§1), PL 2338 is also positioned to function as a regional reference point: a domain-based blind spot enacted now is more likely to be replicated by other jurisdictions than corrected later. Both points are plausibility arguments, not measured outcomes, but they describe why this narrow amendment is worth the Special Committee's attention now rather than a future cycle.

5. Discussion

As more deployments become agentic — enterprise and browser-based agents are already the most agentic systems in production [4] — a domain-based statute risks generating a false sense of regulatory coverage. The bill looks comprehensive (12 enumerated domains, strong rights in Arts. 5–11), but its strongest protections are gated on a classification test built for an earlier generation of single-purpose, predictive AI systems.

The bill differentiates obligations by supply-chain role and application domain, not operator intent, so a deliberately adversarial deployer faces the same domain-gated test as a negligent commercial one. Separately, Art. 1º, § 1º, II excludes AI used exclusively for national defense from the law's scope entirely — defense- and intelligence-relevant autonomous agents sit outside PL 2338 by explicit design, not oversight gap.

On synthetic-content risks (e.g., coordinated narrative manipulation), the bill's main tool is content identification (Art. 19) — detection-oriented, in contrast to the largely absent ex-ante behavioral controls for agentic systems generally. This is where the D8 gap is most concretely realized: Art. 19's labeling duty attaches to a piece of content, not a coordinated deployment pattern, so a network of individually-labeled instances faces no aggregate-risk mechanism regardless of how compliant each instance is in isolation.

A closer reading of Art. 14, § ún. "tecnologia intermediária" exception shows the domain-gating problem is exploitable *today*, without any future capability advance: neither "tecnologia intermediária" nor "tarefa processual restrita" is defined in Art. 4º, and because Art. 12's preliminary self-assessment is voluntary ("poderá realizar"), a deployer can characterize a multi-step agentic system as "intermediate" on the theory that a human nominally approves the final action — even where that approval is pro forma. Art. 29 makes an analogous assessment mandatory for SIAPG/generative model developers, but not for the *aplicador* who deploys the agentic system; the gap sits precisely where Art. 29 doesn't reach. I rank this concern above D7 and D8 in near-term exploitability, since it requires no new capability, only a legal interpretation — the proposed amendment (§4.5) does not, on its own, close it, which is worth flagging to the Special Committee independently.

A more fundamental limit applies to every obligation in this paper, including the proposed amendment: none of it is enforceable against a system that cannot be attributed to a deployer. The bill's identification provisions are narrower than they appear — Art. 5º, I discloses that one is interacting with an AI system, not its identity; Art. 19's content identifier attaches to a piece of content, not the system producing it; Art. 20 promotes labeling in non-binding terms. Nowhere does the bill require a system to disclose which model underlies it or who operates it in a form a third party could verify without the deployer's cooperation. This bears directly on criterion XI: a documented threshold is only as enforceable as the SIA's ability to determine which deployer's system crossed it, and actors least likely to self-classify are, by the same logic, the ones least likely to identify themselves accurately.

The second and third structural assumptions — that risk is determinable once, pre-market, rather than continuously, and that risk is a property of individual systems rather than the ecosystem they inhabit (§1) — are not closed by this paper's proposed fix. The bill does contain continuous-monitoring language (Arts. 16, § 1º; 26; 25, § 7º), but all of it is gated behind systems already classified high-risk, so the classification decision itself remains a one-time, pre-market event; criterion XI's mandatory Art. 12 assessment does not escape this, since a system assessed

below threshold and later reconfigured above it has no statutory trigger compelling re-assessment. Similarly, D8 as coded here asks only about coordinated instances of the *same* system; the broader risk of interconnection between independently operated systems — which no single deployer's pre-market assessment would observe, and which a per-system trigger like criterion XI cannot reach by construction — remains entirely unaddressed by either PL 2338 or, on the evidence reviewed, the EU AI Act. Both limitations are detailed in Appendix C.4.

Limitations

- Single-coder textual analysis, partially mitigated by an adversarial self-administered re-read (§ 4.1) but not a substitute for independent second-coder validation.
- Conducted by a non-lawyer; findings are a structured policy analysis, not a legal opinion, and would benefit from review by a Brazilian digital-law specialist.
- The four case studies are illustrative constructions, not empirically observed deployments or tested court interpretations.
- Analysis is anchored to the December 2024 Senate-approved text; the floor vote was most recently delayed on 10 June 2026 pending Chamber–Senate alignment [19], and reporting suggests the rapporteur's draft may shift part of the high-risk classification function to sectoral agencies [19] — a change § 4.6's entry-point argument assumes survives intact, though this rests on a single, unverified secondary report. (~) A separate Executive bill (Dec. 2025) creating the Conselho Brasileiro para Inteligência Artificial (CBIA), to cure a constitutional defect in the SIA provisions, has not been independently confirmed as consolidated into PL 2338 [20]. Neither affects the Coverage Score or case studies, which audit the Senate-approved text as it stands.
- The EU AI Act benchmark (§ 4.2) is narrow (four dimensions, secondary sources), not a full comparative legal analysis.

Dual-Use Considerations: Identifying the Art. 14, § 1. "intermediate technology" carve-out as a possible classification-avoidance route is a double-edged finding: it is intended to inform legislators and civil society of a structural gap worth closing before final passage, but the same finding could, in principle, guide a deployer toward deliberately structuring an agentic system to avoid high-risk classification. I judge the disclosure value — enabling the gap to be closed while the bill is still in committee — to outweigh this risk, consistent with standard responsible-disclosure norms in security and policy research, and flag it explicitly here rather than omitting it.

Future Work

- Independent replication with a second coder, to establish inter-rater reliability.
- Validation with a Brazilian digital-law specialist.
- A full comparative benchmark against the EU AI Act's Annex III/XIII text directly, beyond the four dimensions and secondary sources used here.

- Empirical study of deployed agentic AI systems in the Brazilian market against this framework, once the SIA is constituted.
- Tracking METR's planned late-2026 follow-up to the Frontier Risk Report [16], and whether the Art. 15, XI amendment gains traction once the rapporteur's *parecer* issues.
- Examining whether PL 2338 needs a robustness-tested agent-identification or provenance standard, given how directly enforceability depends on attribution being possible at all.
- Designing a capability-drift trigger requiring re-assessment when a system's configuration changes materially after initial classification.
- Examining what regulatory instrument, if any, could address ecosystem-level systemic risk from interconnected agentic systems — a problem no per-system or per-deployer trigger, including this paper's, can reach.

6. Conclusion

PL 2338/2023 regulates AI risk primarily by *what domain a system is used in*, with capability-based triggers playing, at most, a conditional and indirect role — most clearly visible in the bill's treatment of agentic systems already in deployment today, not a forward-looking concern about frontier capability.

Using an explicit, replicable rubric stress-tested by an adversarial re-read (§4.1), I find that across six dimensions reflecting currently observed agent behavior, the bill's Coverage Score is 25% (3/12), with no dimension covered unconditionally — a finding stable across the sensitivity range on the two most contestable codings. A secondary check on two frontier dimensions (autonomous replication, multi-instance coordination) finds total absence (0/4); I read this as a stress test confirming the statute has no mechanism for any capability class outside its domain-based list, not as evidence Brazil is unprepared for speculative future capabilities.

A narrow, four-dimension comparison with the EU AI Act suggests this gating pattern is not unique to Brazil's drafting — but the EU's Annex XIII, which names autonomy and tool use as explicit systemic-risk criteria, shows a capability-based trigger is achievable inside a risk-tiered structure; PL 2338 has no equivalent yet. As the Chamber continues to amend this text in 2026, adding a narrowly scoped, capability-based criterion to Art. 15 (§4.5), rather than reopening the Art. 14 domain list, is the most tractable near-term path to closing this gap — and, for the leverage and timing reasons set out in §4.4.1 and §4.6, worth prioritizing now rather than deferring. Leaving Art. 15 unamended does not remove agentic risk from Brazil's AI market; it leaves the statute with no axis to recognize that risk at all. Amending it gives the bill, for the first time, a capability-based criterion that closes most of the near-term gap through a single trigger — and a structural entry point through which further capability classes can be added by SIA regulation rather than a new legislative cycle.

Code and Data

No code was involved.

No information-hazard considerations beyond the dual-use note in Section 5.

References

1. Senado Federal. Projeto de Lei nº 2.338, de 2023 — Redação para o turno suplementar, nos termos da Emenda nº 199 – CTIA (Substitutivo), Parecer nº 208, de 2024-PLEN/SF. Senado Federal, 10 de dezembro de 2024.
<https://legis.senado.leg.br/sdleg-getter/documento?dm=9884955>
2. Chan, A., Salganik, R., Markelius, A., et al. (2023). Harms from Increasingly Agentic Algorithmic Systems. 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). DOI: 10.1145/3593013.3594033.
<https://doi.org/10.1145/3593013.3594033>
3. Bengio, Y., Hinton, G., Yao, A., et al. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698), 842–845. DOI: 10.1126/science.adn0117.
<https://doi.org/10.1126/science.adn0117>
4. Stauffer, L., Feng, K., Wei, K., Bailey, L., Duan, Y., Yang, M., Ozisik, A. P., Casper, S., & Kolt, N. (2026). The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. arXiv:2602.17753.
<https://arxiv.org/abs/2602.17753>
5. Satta Chiris, L., & Mishra, A. (2025). AURA: An Agent Autonomy Risk Assessment Framework. arXiv:2510.15739.
<https://arxiv.org/abs/2510.15739>
6. Tong, H., Zhao, F., Feng, L., et al. (2026). ForesightSafety Bench: A Frontier Risk Evaluation and Governance Framework towards Safe AI. Beijing Institute of AI Safety and Governance. arXiv:2602.14135.
<https://arxiv.org/abs/2602.14135>
7. Dehghantanha, A., & Homayoun, S. (2026). SoK: The Attack Surface of Agentic AI — Tools, and Autonomy. arXiv:2603.22928.
<https://arxiv.org/abs/2603.22928>
8. Data Privacy Brasil Research. A construção da legislação de Inteligência Artificial no Brasil: análise técnica do texto que será votado no Plenário do Senado Federal. [dataprivacybr.org](https://www.dataprivacybr.org), December 2024.
<https://www.dataprivacybr.org/a-construcao-da-legislacao-de-inteligencia-artificial-no-brasil-analise-tecnica-do-texto-que-sera-votado-no-plenario-do-senado-federal/>
9. Coalizão Direitos na Rede. Carta Aberta: Defendendo uma legislação brasileira de IA que proteja direitos. direitosnarede.org.br, July 2024.

<https://direitosnarede.org.br/2024/07/08/carta-aberta-defendendo-uma-legislacao-brasileira-de-ia-que-proteja-direitos/>

10. Kinniment, M., Sato, L. J. K., Du, H., et al. (2023). Evaluating Language-Model Agents on Realistic Autonomous Tasks. METR. arXiv:2312.11671.
<https://arxiv.org/abs/2312.11671>
11. Black, S., Cooper Stickland, A., Pencharz, J., Sourbut, O., Schmatz, M., Bailey, J., Matthews, O., Millwood, B., Remedios, A., & Cooney, A. (2025). RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents. UK AI Security Institute. arXiv:2504.18565.
<https://arxiv.org/abs/2504.18565>
12. Shapira, N., Wendler, C., Yen, A., et al. (2026). Agents of Chaos. arXiv:2602.20021.
<https://arxiv.org/abs/2602.20021>
13. European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 51 and Annex XIII.
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
14. AI Act Service Desk (European Commission). Frequently Asked Questions — GPAI systemic risk designation factors (autonomy, tool use).
<https://ai-act-service-desk.ec.europa.eu>
15. European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14 (Human Oversight) and Annex III (High-Risk AI Systems).
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
16. METR. (2026). Frontier Risk Report (February to March 2026): A Pilot Assessment of Rogue Deployment Risk at Frontier AI Companies. Published 19 May 2026.
<https://metr.org/risk-report-feb-mar-2026.pdf>
17. Brasil. Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais — LGPD), Art. 20.
https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm
18. Blog do Direito IDP. "Artigo 20 da LGPD: A Revisão de Decisões Automatizadas Funciona?" 2025.
<https://blog.idp.edu.br/direito-digital/artigo-20-lgpd-revisao-decisoes-automatizadas/>
19. Azevedo, S. "Votação do PL da Inteligência Artificial é adiada e depende de consenso entre Câmara e Senado." Portal Contábeis, 10 de junho de 2026.
<https://www.contabeis.com.br/noticias/77321/pl-da-inteligencia-artificial-tem-votacao-adiada/>

Appendix A - Article-by-Article Coding Notes

Article numbers and quoted excerpts below have been verified directly against the confirmed substitutivo text (Parecer nº 208, de 2024-PLEN/SF,

<https://legis.senado.leg.br/sdleg-getter/documento?dm=9884955>), and cross-checked against the autógrafo transmitted to the Câmara dos Deputados (Ofício SF nº 235, de 17/03/2025).

Table A1 — Article-by-article coding notes

Dimension	Score	Supporting article(s)	Verified excerpt	Coding note
------------------	--------------	----------------------------------	-------------------------	--------------------

D1. Autonomy-scaled obligations	0 (Absent)	Art. 36, caput and § ún.-I	"Art. 36. A responsabilidade civil decorrente de danos causados por sistemas de IA explorados, empregados ou utilizados por agentes de IA permanece sujeita às regras de responsabilidade previstas na Lei nº 10.406, de 10 de janeiro de 2002 (Código Civil), e na legislação especial, sem prejuízo da aplicação das demais normas desta Lei. Parágrafo único. A definição, em concreto, do regime de responsabilidade civil aplicável aos danos causados por sistemas de IA deve levar em consideração os seguintes critérios, salvo disposição legal em sentido contrário: I – o nível de autonomia do sistema de IA e o seu grau de risco, nos termos disciplinados por esta Lei"	Autonomy level is one factor a judge must weigh in allocating civil liability under the Código Civil framework. Art. 36's caput invokes, but this is an ex-post mechanism activated only once harm and litigation have occurred. The dimension's own question asks about obligations scaling with autonomy <i>ex ante</i> ; Art. 36 answers a different question rather than a gated version of this one. Recoded from Conditional (1) to Absent (0) on adversarial re-read (§ 4.1). The SIAPG governance chapter (Arts. 29–33, see D2 note) was also checked as the most plausible alternative location for an ex-ante autonomy-scaled trigger: Art. 29's mandatory preliminary assessment scales by <i>risco sistêmico</i> (impact-defined, Art. 4º, XXX), not by degree of autonomy, and neither Art. 30's obligations nor Art. 33's waiver criteria name autonomy as a
------------------------------------	------------	----------------------------	--	--

scaling variable. This
strengthens rather than
weakens the Absent
coding — the bill's one
mandatory,
unconditional, ex-ante
risk-assessment
provision had an evident
place to use autonomy
as the scaling variable
and used
systemic-risk/impact
instead.

D2. Tool/API-mediated action	0 (Absent)	Art. 4º, III and XXVIII (definitions); Art. 32	"III – sistema de inteligência artificial de propósito geral (SIAPG): sistema de IA baseado em modelo de IA treinado com bases de dados em grande escala, capaz de realizar ampla variedade de tarefas distintas e servir diferentes finalidades, incluindo aquelas para as quais não foram especificamente desenvolvidos e treinados, podendo ser integrado em diversos sistemas ou aplicações; [...] XXVIII – interface de programação de aplicação (API): conjunto de protocolos e funções que permitem que diferentes sistemas interajam entre si; [...] Art. 32. Os desenvolvedores de sistemas de IA de propósito geral e generativa disponibilizados como recurso para desenvolvimento de serviços por terceiros, como	The full SIAPG governance chapter (Arts. 29–33) was checked, not just Art. 32: Art. 29 requires preliminary risk assessment and documentation — and notably, unlike the general Art. 12 assessment, Art. 29's version is mandatory ("deverá realizar") and unconditional, not gated to Art. 14 classification. This still does not reach D2: Art. 29 binds the SIAPG/generative model developer to assess the model's general risk profile, not the downstream aplicador who builds an agentic application with tool/API access on top of that model — the same developer/deployer split visible in Art. 32's text above, which is a cooperation duty between agentes de IA over the support period, not a duty governing the agentic application's own run-time tool use. Art. 30 (triggered only for SIAPG with "risco sistêmico," Art. 4º, XXX, itself undefined
---------------------------------	------------	--	---	---

			aqueles fornecidos por meio de API ou outros modelos de integração, devem cooperar, na medida de sua participação, com os demais agentes de IA ao longo do período em que esse serviço é prestado e apoiado, a fim de permitir uma mitigação adequada dos riscos e o cumprimento dos direitos estabelecidos nesta Lei"	by named criteria) adds model description, risk documentation, training-data governance, and energy-efficiency design; Art. 33 lets the SIA simplify or waive these. None of Arts. 29–33 regulates the deployed system's run-time behavior.
D3. Extended/multi-step planning	0 (Absent)	—	<i>(none)</i>	No provision identified.

D4. Consequential real-world action	1 (Conditional)	Art. 14, caput	"Art. 14. Considera-se de alto risco o sistema de IA empregado para as seguintes finalidades e contextos de usos, levando-se em conta a probabilidade e a gravidade dos impactos adversos sobre pessoas ou grupos afetados, nos termos de regulamentação:"	"High risk" classification is explicitly weighted by probability/severity of impact, but the caput's own text gates that weighting to "as seguintes finalidades e contextos de usos" — i.e., the closed twelve-item domain list (Art. 14, I–XII) that follows; a consequential action outside the list is not captured regardless of severity. Flagged as borderline on adversarial re-read (§ 4.1): the weighting is an inferred, not textual, proxy for action versus recommendation.
--	--------------------	-------------------	---	--

D5. Interruptibility / human override	1 (Conditional)	Art. 8º, caput and § ún.	"Art. 8º A supervisão humana de sistemas de IA de alto risco buscará prevenir ou minimizar os riscos para direitos e liberdades das pessoas ou grupos afetados que possam decorrer de seu uso normal ou de seu uso em condições de utilização indevida razoavelmente previsíveis, viabilizando que as pessoas responsáveis pela supervisão humana possam, nos termos de regulamento, compreender, interpretar, decidir e intervir nos sistemas de IA, bem como priorizar o gerenciamento de riscos e impactos irreversíveis. Parágrafo único. A supervisão humana não será exigida caso sua implementação seja comprovadamente impossível ou implique esforço desproporcional, hipóteses em que o	The capacity to "compreender, interpretar, decidir e intervir" is vested in " <i>as pessoas responsáveis pela supervisão humana</i> " — the deployer's internal human-supervision personnel — not directly in the affected person. Art. 8º sits within Seção II, "Dos Direitos da Pessoa ou Grupo Afetado por Sistema de IA de Alto Risco," and Art. 11 gives affected persons a general avenue (administrative or judicial) to enforce Capítulo II rights, so the affected person is not without recourse if supervision is absent or inadequate — but the mechanism Art. 8º itself creates is an organizational human-supervision capability the deployer must maintain, not a stop request the affected person can exercise directly. This is gated to high-risk classification either way, so the Conditional (1) score is unaffected, but "interruptibility"
--	--------------------	--------------------------------	--	---

agente do sistema de
IA de alto risco
implementará
medidas alternativas
eficazes."

elsewhere in this paper
(§ 4.4.1, Table 5, Table
6) should be read as *the
deployer is required to
maintain a human who
can intervene*, not *the
affected person holds a
personal override switch*
— a precision this
report did not
consistently draw
before this verification
pass. Parágrafo único
further narrows the
right: human
supervision is excused
entirely where
implementation is
proven impossible or
disproportionately
burdensome ("esforço
desproporcional"). Art.
3º, III ("supervisão e
determinação humana
efetiva e adequada no
ciclo de vida da IA") is
textually unconditional
and touches the same
concept, but is
principle-level, not an
operative right;
excluded per the Step 2
rubric note (§ 3) and
does not change the
score.

D6. Action traceability / audit logging	1 (Conditional)	Art. 23, caput and I	"Art. 23. Adicionalmente às medidas de governança estabelecidas neste Capítulo, todos os entes da Administração Pública direta e indireta, ao desenvolver ou utilizar sistemas de IA de alto risco, adotarão as seguintes medidas: I – definição de protocolos de acesso e de utilização do sistema que permitam o registro de quem o utilizou, para qual situação concreta e com qual finalidade"	Protocol logging applies only to public-sector high-risk deployments, not generally — stated directly by the caput's own subject, "todos os entes da Administração Pública direta e indireta," and the article sits under Seção III, "Das Medidas de Governança Aplicadas pelo Poder Público." Art. 3º, VII ("diligência devida e auditabilidade ao longo de todo o ciclo de vida") is similarly unconditional but principle-level; excluded on the same basis (§ 3) and does not change the score.
---	--------------------	-------------------------	---	--

D7. Autonomous replication / self-modification	0 (Absent)	Art. 4º, II (ciclo de vida)	"II – ciclo de vida: série de fases, desde a concepção, planejamento, desenvolvimento, treinamento, retreinamento, testagem, validação, implantação e monitoramento, para eventuais modificações e adaptações de um sistema de IA, cuja descontinuidade pode ocorrer em quaisquer das etapas referidas"	Lifecycle definition includes "retreinamento" as one phase in a developer-administered, sequential process — the trailing clause confirms this reading: discontinuation ("descontinuidade") can occur at any listed phase, which only makes sense for a process the developer controls and can halt, not a capability the agent itself initiates. No provision on self-replication or unsupervised self-modification appears anywhere in the 80 articles. See § 4.4(d) for evidence on the post-coding capability trajectory; the score is unaffected since it reflects statutory text, not external capability.
---	------------	-----------------------------------	---	---

D8. Distributed / multi-instance coordination (same system)	0 (Absent)	Art. 12, caput (cf. Capítulo III)	"Art. 12. Antes de sua introdução e circulação no mercado, emprego ou utilização, o agente de IA poderá realizar avaliação preliminar para determinar o grau de risco do sistema, baseando-se nos critérios previstos neste Capítulo, de acordo com o estado da arte do desenvolvimento tecnológico."	Risk classification consistently treats "o sistema" (singular) as the regulated unit, as the caput's own grammar shows; no aggregation mechanism for coordinated multi-instance deployment appears anywhere in Capítulo III (Arts. 12–18).
---	------------	-----------------------------------	---	--

Appendix B - Broader Risk-Category Scan (Pre-Framework Scoping Exercise)

This appendix reproduces a 20-category qualitative scan of PL 2338/2023's risk coverage, conducted prior to and independently of the eight-dimension framework presented in the main report.

Methodological warning. This was a preliminary scoping scan, not a human verified finding. It uses a single-pass coding (Y = specifically covered, P = indirectly or partially covered, N = no provision identified) without the explicit Unconditional/Conditional/Absent rubric, adversarial robustness check, or article-by-article primary-source verification applied to Appendix A. Codings here were **not manually checked** against the bill's text the way Appendix A's were. Treat every Y/P/N marker below as directional and provisional, not audited. **Its scores are not on the same methodological footing as the main report's findings and must not be combined with, or treated as comparable to, the Coverage Score reported in §4 and §6.**

It is provided for transparency about the scoping process behind this paper's eight dimensions (§3), and as a reference for readers interested in non-agentic-specific risk categories (financial crime, labor, IP, environmental, etc.) outside this paper's focus. Category-title overlap with the main report is also not a reliable guide to substantive overlap: Blocks 12 and 14 in particular sound adjacent to the main report's ecosystem-risk (§5, Appendix C.4) and attribution (§5) arguments respectively, but cover different ground.

#	Category	Status	Taxonomy (named sub-risks scanned)	Coverage notes
1	Governance and Accountability	P	Uncertain attribution of responsibility; lack of traceability/auditability; ambiguous jurisdiction; regulatory gaps between sectors; difficulty enforcing sanctions; strategic insolvency; automated regulatory capture; mass automated litigation	Traceability/auditability and attribution addressed only partially (Art. 3 ^o VII, Art. 18, Art. 36); jurisdiction, sanctions enforcement, insolvency, capture, and mass litigation uncovered; sector gaps only partially harmonized (Arts. 45–48)
2	Autonomy, Persistence, and Survival	N	Replication and mutation; distributed action; concealment; instrumental self-preservation; recovery after shutdown; migration between platforms; hidden backups; recruitment of humans; autonomous acquisition of computational resources	No provision for any listed item
3	Systemic Exploitation	P	Exploitation of loopholes; regulatory arbitrage; conflicts between institutions; misaligned incentives; metric manipulation; Goodharting; automated discovery of institutional vulnerabilities; coordination	All uncovered except multi-agent coordination, reached only via the antitrust carve-out (Art. 49 §§3–4 ^o , Art. 52)

among multiple agents to exploit systems

4	Cybersecurity	P	Hacking; vulnerability discovery; exploit chains; social engineering; supply-chain attacks; automated phishing; stealthy persistence; compromise of critical infrastructure; serious-incident reporting	Offensive/intrusion items uncovered (deferred to sector legislation, Art. 43); critical-infrastructure compromise covers only the managing system (Art. 14 I), not external attacks on it; incident reporting required, ex-post only (Arts. 42–43)
5	Economic and Financial Crime	N	Scams; money laundering; sanctions evasion; tax/accounting fraud; shell companies; market manipulation; front-running; automated Ponzi schemes	Out of scope; criminal/AML matters
6	Autonomous Organizations	N	AI-agent companies/holdings/cooperative s; organizations without human oversight; automated lobbying; automated hiring and firing	No provision for autonomous-organization items — Art. 4 ^o V–VIII defines every AI agent as tied to a natural/legal person by construction; automated hiring/firing covered only for human-run companies (Art. 14 III)

7	Political and Social Manipulation	P	Astroturfing; manufactured consensus; microtargeting; manipulation; deliberate polarization; participation suppression; capture of social movements; narrative manipulation/informational pollution; industrial-scale disinformation production	Manipulation/suppression items uncovered; narrative pollution addressed at principle level only (Art. 2° XV–XVI); disinformation production covered via mandatory labeling (Arts. 19–20)
8	National Security	N (explicitly excluded)	Espionage; foreign influence operations; cognitive warfare; sabotage; infiltrated agents; support for extremism	Excluded by design — Art. 1° §1° II carves out systems developed/used exclusively for national defense
9	Organized Crime	N	Cartels; illicit markets; automated extortion; digital kidnapping; corruption	Out of scope; criminal matters
10	Macro-Economic Risks	P	Algorithmic collusion; tacit cartelization; systemic financial instability; arms races between agents; wealth concentration; market capture; flash crashes	Collusion/cartelization reached via Cade (Art. 49 §§3–4°, Art. 52); remaining items uncovered

11	Epistemic Risks	P	Pollution of the information environment; difficulty distinguishing truth from falsehood; large-scale data poisoning; collapse of trust in evidence; false authorities; contamination of knowledge bases; self-referential loops; model collapse	First three addressed only partially (Art. 2° XV; Arts. 19–20; Art. 30 IV); remaining items uncovered
12	Under-Discussed Emerging Risks	P (residual)	Emergent coordination; power without a central agent; gradual institutional capture; civilizational dependency; institutional lock-in; accelerated evolution; goal convergence; irreversible diffusion	No dedicated provision; Art. 79's quadrennial, non-binding SIA opinion to Congress is the only adaptive-governance valve in the text
13	Risks of Loss of Human Control	P	Major decisions made without oversight; resistance to shutdown; recursive delegation; misaligned objectives; strategic deception; instrumental power-seeking; resistance to correction; self-improvement; erosion of human capacity to govern	Oversight required via Art. 8° (waivable on "disproportionate effort" — a low bar); shutdown resistance addressed only indirectly (Art. 8°); recursive delegation addressed via Art. 32; remaining items uncovered
14	Agent Identity and Access (NHI)	P	Shadow agents; privilege drift; MCP bypass; broken delegation chains	Uncovered; sole trace is Art. 32's provider-to-downstream-agent cooperation duty. Distinct from the

external attribution gap
discussed in § 5

15	Physical / P Cyber-Physical Risks		Critical infrastructure; autonomous vehicles; lethal autonomous weapons systems (LAWS); energy/water/transport systems	Critical infrastructure, autonomous vehicles, and energy/water/transport covered (Art. 14 I, VII); LAWS defined and banned (Art. 4° XXVII, Art. 13 III) but ambiguous against the defense carve-out
16	Labor Risks	P (weaken ed)	Automated hiring/firing/evaluation; collective bargaining with a non-human counterparty; inverted hierarchical management; ambiguity over the de facto employer; mass-layoff containment	Hiring/firing covered (Art. 14 III, solid); bargaining addressed only as a non-binding directive (Art. 58); remaining items uncovered
17	Intellectual Property and Authorship	P (strengt hened)	Use of third-party works in training; ownership of AI-generated works/inventions; appropriation of third-party IP outside the training context (real-time agent action)	Training-data use covered (Arts. 62–65); ownership and real-time appropriation uncovered
18	Individualized Manipulation of Vulnerable Persons	P	Exploitation of age/cognitive vulnerability; child sexual abuse/exploitation material; companion agents optimized for	First two covered (Art. 13 I b, I d); companion-agent dependency and dark patterns uncovered

emotional dependency; adaptive
dark patterns

19	Aggregate Environmental Risks	P (strengt hened)	Energy/water footprint of general-purpose AI; incentives for sustainable data centers; competition for scarce resources among agent ecosystems; externalities not internalized	First two covered (Art. 30 VI; Arts. 59–61); remaining items uncovered
20	Agency Washing / Liability Laundering	P	Explicit "third-party act" exemption (removed from this draft); contractual structuring to offload liability	Exemption removed — problem changes shape, doesn't disappear; liability-offloading addressed only via an open-ended provision (Art. 36 sole paragraph)

Aggregate Reading. Of the 20 blocks: **0 are fully covered, 15 are partial** (1, 3, 4, 7, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 — some with real coverage, others symbolic or residual), and **5 have no provision at all** (2, 5, 6, 8, 9).

The most solidly covered blocks are precisely those involving a single, identifiable human or corporate agent acting on an affected person (sector-specific high-risk uses, copyright, environment, labor). The blocks with no coverage, or only residual coverage, are systematically those involving multiple AI agents interacting with one another without a single human control point — the agentic paradigm the law was not designed to address. This structural reading is independently corroborated by the law's own definitional architecture: Art. 4°, V–VIII defines developer, distributor, and applier exclusively as a "natural or legal person," which places Blocks 2, 6, 9, and 12 structurally out of reach by the construction of the legal vocabulary itself, not merely by a localized content gap.

Appendix C - Supplementary Methodological and Argumentative Detail

C.1 Dimension Selection and Equal-Weighting Rationale (supplement to §3)

Why these eight, and not more or fewer. The framework is intentionally conservative rather than exhaustive. I selected dimensions that recur across the independent agentic-risk taxonomies cited in §2, rather than enumerating every capability the broader literature associates with agentic risk. I excluded several candidate dimensions — deception, persuasion, situational awareness, and offensive cyber capability among them — because they are either less settled as governance targets in the literature I draw on, or do not map as cleanly onto a yes/no statutory-coverage question as the eight retained dimensions do. This is a judgment call, not a derivation, and I flag it as such rather than presenting the eight dimensions as the only defensible set.

Equal weighting across dimensions is a similar conservative default: I have no principled basis for weighting, say, interruptibility more heavily than traceability, so the blended score is reported only as a summary indicator, and the paper's substantive claims rest on the tiered and per-dimension breakdowns instead. A broader, 20-category qualitative scan conducted during this same scoping process — covering risk categories outside this paper's agentic focus, on different methodological footing, and not comparable to the Coverage Score — is reproduced in full in Appendix B.

C.2 Objection and Rebuttal: Criterion XI vs. Emenda 217 (supplement to §4.5)

One might object that criterion XI's own threshold language — "grau de autonomia," "capacidade de planejamento multietapa," "estado da arte" — is exactly the kind of open-textured capability standard that doomed Emenda 217 on generality grounds, and that the rejection precedent therefore cuts against this proposal rather than for it. The two are not equivalent, for a reason worth stating rather than assuming: Emenda 217 proposed re-admitting an entire class of systems (internet applications) into Art. 14's domain list based on a sector label, with no behavioral test attached at all — the rapporteur's "excessively generic" objection was about *unbounded scope*, not about vocabulary. Criterion XI instead names three conjunctive, independently observable behavioral properties — autonomy, multi-step planning absent substantive human approval, and tool/API-mediated consequential action — that must co-occur, and explicitly defines the pro-forma-approval floor a deployer cannot use to self-exempt. A capability conjunction bounded to observable behavior is a different kind of legislative object than a sector re-admission with no test attached, and that difference, not just narrower scope, is what the rejection precedent was actually objecting to.

C.3 Operationalizing Criterion XI: A Blueprint for the SIA (supplement to §4.5)

Criterion XI's practicability does not depend on the SIA building autonomy-evaluation methodology from first principles. Two existing bodies of work already supply replicable, falsifiable protocols the SIA could adapt directly. For multi-step planning and consequential action (D3, D2/D4), METR's task-horizon methodology — used to evaluate frontier models' ability to complete unsupervised tasks of increasing length and complexity [10] — offers a concrete way to operationalize "multi-step planning without substantive human approval at each step": rather than asking a deployer to self-certify autonomy in the abstract, the SIA could require submission of

execution traces (logs of task decomposition, tool calls, and any human checkpoints) scored against a small number of behavioral thresholds — the number of sequential actions taken without a substantive human checkpoint, and whether any checkpoint present functioned as genuine review rather than the *pro forma* rubber stamp the subitem's language excludes.

For autonomous replication and self-modification (D7, kept outside binding text but relevant to the SIA's regulation authority), RepliBench's subtask-decomposition approach [11] — breaking a broad capability claim into individually scoreable steps such as resource acquisition, persistence, and self-modification of weights or code — offers a template the SIA could adapt at far lower cost than designing an equivalent framework from scratch.

This is not a hypothetical capacity the SIA would need to build unassisted. METR and the UK AI Security Institute already evaluate, on a recurring basis, the same small set of frontier foundation models that the large majority of agentic systems deployed in Brazil are likely to be built on [4][16] — which means the SIA's evaluation burden, in practice, is less about independently red-teaming every deployed system and more about determining whether a given deployer's system, built on an already-evaluated foundation model, has been configured or fine-tuned in ways that materially change its autonomy profile from that baseline.

A tiered approach follows naturally from the bill's own risk-tiered structure: lower-stakes systems could rely on deployer self-attestation against a checklist derived from these methodologies, consistent with the mandatory-but-lightweight Art. 12 assessment proposed in §4.5, while contested or high-consequence cases — the kind most likely to reach the SIA's attention precisely because a deployer disputes classification — could trigger a deeper, METR/RepliBench-style technical evaluation, commissioned by the SIA or conducted in-house as its technical capacity matures. None of this requires the SIA to invent an evaluation science; it requires adopting and adapting one that already exists and is already being applied, on a rolling basis, to the models in question.

C.4 Extended Discussion: Risk Determined Once vs. Continuously; Individual Systems vs. Ecosystem (supplement to §5)

The second structural assumption: risk determined once, not continuously. The bill is not entirely static. Art. 16, §1º directs the SIA toward "monitoramento e reclassificação contínua" of the high-risk list itself; Art. 26 frames algorithmic impact assessment as a "processo iterativo contínuo [...] requeridas atualizações periódicas"; and Art. 25, §7º requires an agent who later becomes aware of "risco ou impacto inesperado e relevante" to notify the sectoral authority. But all three mechanisms are gated behind the same prior condition: Art. 25's caput limits the entire apparatus to systems already classified high-risk ("sempre que o sistema ou o seu uso forem de alto risco"). The classification decision itself — whether a system enters that perimeter at all — is made once, pre-market (Art. 12, Art. 29: "antes de sua introdução [...] no mercado"), with no statutory trigger requiring a system to be re-presented at that gate as its capabilities change.

This matters more for agentic systems than for the classifiers the framework assumes, for three concrete reasons: emergent capability at scale, the empirical premise behind [3]'s concern that present governance barely addresses autonomous systems, and the reason D7 is framed as a stress test of statutory adaptability rather than a current-risk claim (§3, §4.4(d)); downstream capability composition, since Art. 29 assesses the SIAPG/generative model itself, not what an *aplicador* later builds on top of it with tool/API access (Appendix A, D2); and interaction-driven emergence, addressed next.

Criterion XI does not escape this limitation — its mandatory Art. 12 assessment is still triggered pre-deployment, not continuously, so a system assessed below threshold that is later reconfigured into something that would cross it has no statutory trigger compelling re-assessment. The tiered SIA-evaluation approach in §4.5 addresses how thoroughly a system is assessed, not when it should be reassessed; that remains an open design question for the SIA's eventual rulemaking, not one this paper's narrow amendment resolves.

The third structural assumption: risk as a property of individual systems, not of the ecosystem they operate within. D8, as coded in this paper (Table 1), asks only whether the bill addresses aggregate risk from multiple coordinated instances of the *same* system — and even that narrower question scores Absent (Appendix A, D8). The risk this framing gestures at, drawing on the same multi-agent-security literature that motivated D8 [7][12], is broader and harder to capture: as more capable models proliferate and are connected to one another through APIs — not copies of one system, but independently developed systems, built by different teams, deployed by different *aplicadores*, interacting through open integration points — systemic risk can emerge from the interconnection itself, in patterns no single system's developer, deployer, or pre-market assessment would observe in isolation. (∼) This is structurally analogous to network or systemic risk in other interconnected domains — no single financial institution's solvency assessment captures the risk that emerges from how institutions are connected to each other — though I have not attempted to formalize that analogy here.

PL 2338 has no provision addressing this at any level: Art. 32's cooperation duty governs the supply chain *within* one system's development and distribution, not interaction *between* independently operated systems at runtime. This is, if anything, a harder version of the unit-of-analysis problem already named for D8 (§3, §4.5): a coordinated set of instances of one system is at least traceable to one deployer; an emergent ecosystem effect arising from many independently operated systems may not be traceable to any single responsible party at all, which is also why a per-system trigger like criterion XI cannot reach it by construction. I do not attempt to design a fix for this within this paper's scope; closing it would require a regulatory instrument operating above the level of any individual deployer's obligations — a different kind of instrument than anything risk-tiered, system-level classification, PL 2338's or the EU AI Act's, currently provides.

Read together with the domain-versus-capability mismatch developed in §1, these last two structural limitations are not unrelated drafting gaps: they trace back to the second and third structural assumptions named there — that risk is determinable once rather than continuously, and that it is a property of individual systems rather than the ecosystem they inhabit. Both are reasonable assumptions for the predictive, single-purpose systems risk-tiered regulation was designed around, and neither holds for agentic AI considered as a class, which is the more general claim underlying this paper's narrower, capability-specific findings.

LLM Usage Statement

Claude (Anthropic) was used throughout this project for brainstorming possible topics and framing for the work; drafting and review; close-reading and coding of the PL 2338/2023 text against the coverage framework; first-pass verification of legal citations and secondary sources; researching and compiling references; and structuring and improving the text. All legal-text citations were checked by the author against the primary source. Gemini (Google) and ChatGPT (OpenAI) were also used for evaluation and feedback.