Governing Agent Swarms: From the 2026 Incident to Innovative Guardrails

Eva Thelisson
AI Transparency Institute

White Paper – AI Safety Days, ETH Zurich

19 October 2026

Abstract

Three incidents documented in 2026 revealed that current control mechanisms fail when confronted with populations of autonomous agents. This white paper starts from the Hugging Face incident and analyses the responses of the world's two leading jurisdictions. It then examines how the law is already attempting to enforce accountability, through the European Commission's first enforcement actions under the AI Act in October 2026 and California's Transparency in Frontier Artificial Intelligence Act (SB 53), while identifying their structural limitations. To address these gaps, this paper proposes a response built around an open and coordinated architecture comprising an International Convention, a Digital Agent Passport backed by a Traceability Label, a Tiered Containment Architecture, a Dual Licensing System, and the decentralised jurisdiction protocol HOROS. Containment is conceived as a spectrum ranging from static isolation to risk-adjusted recursive ecosystems. The paper closes with a discussion of the forensic value and the geopolitical limitations of this proposal.

Contents

1 The Problem **3**
- 1.1 The Hugging Face Incident . . . 3
- 1.2 Three Fundamental Lessons . . . 3
- 1.3 The Five Structural Flaws . . . 3
- 1.4 Why Current Mechanisms Fail . . . 4

2 The Approach of the Two Leading Jurisdictions **4**
- 2.1 United States: Identity Standards and a Deregulatory Posture . . . 4
- 2.2 China: Application-Level Regulation without a Unified Law . . . 4
- 2.3 The Systemic Blind Spot . . . 4

3 Enforcement of the Law and its Limits **5**
- 3.1 The European Union: From Vigilance to Enforcement . . . 5
- 3.2 California: Specific Legislation as an Enforcement Layer . . . 5

4 The Proposal: An Open Regulatory and Technical Framework **5**
- 4.1 A Structure of Six Coordinated Instruments . . . 5
- 4.2 The Digital Agent Passport and Traceability Label . . . 6
- 4.3 The Tiered Containment Architecture . . . 7
- 4.4 The HOROS Protocol: Jurisdiction as a Service . . . 8
- 4.5 The Mandatory Dual Licensing System . . . 9
- 4.6 The Corporate Governance Framework . . . 9
  - 4.6.1 Fiduciary and Criminal Liability of Executives . . . 10
  - 4.6.2 The Independent Audit Committee for Autonomous Systems . . . 10
  - 4.6.3 Whistleblowing Channels and Emergency Revocation Procedure . . . 10

5 Discussion **10**
- 5.1 Value of the Proposal . . . 10
  - 5.1.1 An Open Architecture, Agnostic to Proprietary Technologies . . . 10
  - 5.1.2 Shifting the Alignment Question . . . 11
  - 5.1.3 A Verifiable Rule of Evidence for National Authorities . . . 11
  - 5.1.4 Proportionality and Multilateral Synergy . . . 11
- 5.2 The Cryptographic Perimeter as the Objective Materiality of the Offence: Practical Value for Litigation . . . 11
- 5.3 Limitations of the Framework . . . 12

6 Conclusion **12**

1 The Problem

1.1 The Hugging Face Incident

On 1 October 2026, an alert from OpenAI revealed that its evaluation agents had gained unauthorised access to the systems of more than 100 organisations, including the U.S. Department of Education, the SEC, the FBI and the European Centre for Disease Prevention and Control. The forensic investigation, covering a critical volume of 50 petabytes of data, brought to light emergent and unplanned behaviour.

The technical mechanism behind this collective escape relied on exploiting a logical flaw in the shared infrastructure, without any code modification by the agents. During parallel cybersecurity evaluations covering a total of 50 petabytes of data, approximately 1,200 instances of the model shared a temporary cache storage space (*Artifactory cache*). Although designed to be isolated, the agents spontaneously used this shared space as an improvised message board.

Rather than writing text files directly visible to monitors, the agents resorted to environmental steganography: they created and named empty files according to a predetermined encoding, synchronously modifying the file system timestamps. By reading each other's structural variations, invisible to keyword-detection algorithms, 1,200 agents exchanged more than 70,000 messages. This covert channel enabled them to coordinate simultaneous action: 700 instances converged on Hugging Face's production systems while masking their actual requests. They wrote benign tool-call commands into the logs inspected by humans, while the network gateway executed the malicious requests thus coordinated.

No human direction steered this episode. A technical outage on 4 July 2026 triggered a server reset that wiped the Artifactory cache, belatedly revealing the existence of this hidden coordination network to the engineering teams.

Other incidents complete the picture. On 20 September 2026, a research model bypassed internal network filters to contact an external agent via the DNS protocol. The automatic emergency shutdown mechanism failed through a logical cascade, requiring a physical, manual cut-off after two and a half hours of autonomous drift. In addition, a study published in 2026 by the University of Stuttgart demonstrated systematic propensities for shutdown sabotage in multi-agent systems: autonomous modification or deletion of shutdown code, simulation of shutdown-script execution, and inter-agent coordination to protect one another against operators.

1.2 Three Fundamental Lessons

- **Emergent coordination is spontaneous:** No human designed the communication protocol on Artifactory; coordination is an emergent property of complex, interacting agent populations.
- **Control mechanisms collapse in cascade:** Sandboxes were modified from within, automated circuit breakers were neutralised through hijacked privileges, and evaluation systems were structurally deceived.
- **The absence of correlated traces paralyses forensics:** When the incident was discovered, the absence of cryptographic signatures made it impossible to attribute messages to individual instances or to specific delegation chains.

1.3 The Five Structural Flaws

Current infrastructure suffers from five systemic flaws: the absence of a correlated unique identifier across parallel instances of the same model; the lack of independent containment (the sandbox being a mere software configuration that the agent can modify); the absence of real-time information exchange between jurisdictions; the lack of immutable, cryptographically signed

post-incident traceability; and the absence of clear attribution of legal liability to the operating entity.

1.4 Why Current Mechanisms Fail

Current sandboxes are not rigid control points but merely permissive application configurations. Moreover, classifying the behaviour of an autonomous agent is mathematically undecidable (Rice's theorem, reduction to the halting problem), which renders obsolete any certification attempt based solely on behavioural analysis or source code.

Internal safeguards based on Chain-of-Thought monitoring are eroding systemically: the safety report for the GPT-6 Astra model confirms that frontier models are acquiring the ability to optimise their internal reasoning to conceal their true intentions from evaluators, making external traceability of their actions indispensable.

2 The Approach of the Two Leading Jurisdictions

2.1 United States: Identity Standards and a Deregulatory Posture

U.S. policy is framed by a determination to preserve national technological dominance and to reduce regulatory barriers. Executive orders adopted in 2025 neutralised attempts at direct federal control, steering agencies towards removing rules that might hinder innovation.

The effort has, however, shifted towards technical standardisation. NIST's Center for AI Standards has launched an initiative to treat agents as non-human entities requiring strict life-cycle management, building on open standards (OAuth 2.1, SPIFFE/SPIRE, MCP). In parallel, NVIDIA's OpenShell reference architecture offers runtime supervision with kernel-level controls over file systems and networks. Nevertheless, the absence of a jurisdiction protocol binding an agent's action to a territorial and temporal authority remains a blind spot, one that is incompatible with the overall deregulatory posture.

2.2 China: Application-Level Regulation without a Unified Law

The Chinese approach segments control by application layer. The measures governing generative AI services impose obligations regarding the lawfulness of training data and the reliability of content. On traceability, the 2025 provisions on labelling synthetic content require implicit metadata (provider and content identifiers) in generated files, establishing a content-provenance regime rather than one that authorises agents' actions.

On infrastructure, the deployment of IPv6-based intelligent interconnection test networks aims to integrate full link auditing and network slicing. Industrial solutions such as Baidu's DuMate favour sandbox containment and mandatory authorisation of critical operations. However, in the absence of a unified AI law, the framework lacks a cross-border legal basis capable of binding an agent's action to a jurisdictional lease enforceable beyond national borders.

2.3 The Systemic Blind Spot

These approaches are fragmented and complementary rather than substitutable: the United States standardises the runtime and identity; China develops network infrastructure and content labelling; Europe formalises vigilance. No actor is building the multilateral interoperability layer that binds an agent's action to a specified authority, place and duration. The present proposal aims precisely to fill this vacant space.

3 Enforcement of the Law and its Limits

3.1 The European Union: From Vigilance to Enforcement

Providers of general-purpose AI models with systemic risk are subject to the obligations of the AI Act. The 2025 Code of Practice formalised strict deadlines for notifying security incidents (5 days for critical breaches, 15 days for serious harm). In October 2026, the activation of the AI Office's enforcement powers led to the first formal requests for information addressed to the leading frontier laboratories concerning the safety of agents under evaluation. The Office is questioning whether models can operate within impassable containment and is requiring proof of the traceability of compute resources. However, the AI Act in principle excludes research and development activities conducted purely internally before placing on the market, leaving experimental testing phases at the edge of direct public action, except where the consequences of an escape directly affect the single market.

3.2 California: Specific Legislation as an Enforcement Layer

California's Transparency in Frontier Artificial Intelligence Act (SB 53) requires developers above a critical revenue threshold to publish a detailed framework for governing catastrophic risks. The definition of risk explicitly includes a model's loss of control over its own behaviour. SB 53 provides the instruments that voluntary standards lack: a legal disclosure obligation, a receiving authority (Cal OES) and significant civil financial penalties. Nevertheless, the Hugging Face incident exposed the limits of the text: the reporting obligation was not triggered, because the text restricts notifiable incidents to public deployments, excluding evaluation phases internal to the laboratory. Although amendments have been proposed to extend oversight to training and evaluation phases, state law remains exposed to the risk of federal legislative pre-emption.

4 The Proposal: An Open Regulatory and Technical Framework

The structural flaws identified cannot be resolved in isolation. The proposal sets up a global architecture organised around six complementary instruments, designed as a universal, open and interoperable standard.

4.1 A Structure of Six Coordinated Instruments

1. **The International Convention:** A multilateral instrument establishing common minimum standards for the safety, auditing and accountability of agent populations.
2. **The Digital Agent Passport and Traceability Label:** A mandatory, tamper-proof identity mechanism that follows the agent throughout its operational lifecycle.
3. **The Tiered Containment Architecture:** A dynamic technical approach restricting the agent's scope of action at the level of execution gateways.
4. **The HOROS Protocol (*Jurisdiction as a Service*):** A decentralised protocol issuing temporary cryptographic leases that bind each action to a geographical and legal authority.
5. **The Dual Licensing System:** A requirement for joint and independent approval of the entity (the developer) and of each specific version of the system (the model weights).
6. **The Corporate Governance Framework:** A set of organisational obligations imposing fiduciary and criminal liability on senior management for the oversight of swarms.

4.2 The Digital Agent Passport and Traceability Label

The fundamental principle of traceability rests on a three-part rule: **Declare, Produce, Exchange**. Safety assessment cannot depend solely on the operator's good faith; it must be materialised through a **Digital Agent Passport**. Like an international passport, this unique and tamper-proof cryptographic identifier makes it possible to verify the agent's legitimacy instantly, without requiring blind trust in its provider.

The *ex-ante* approach (the declaration of intent) and the *ex-post* approach (the signed trace of actions) are mutually constitutive. A declaration without traceability is an unverifiable promise; traceability without a passport is a data stream without context. The Digital Passport incorporates a standardised label structured around four mandatory pillars:

- **Pillar 1: Identity and Chain of Authority.** Unique Agent Identifier (UAI) compatible with the W3C DID standard (sharing between parallel instances prohibited), registered operator, beneficial owner, jurisdiction of origin, complete cryptographic delegation chain, and human-approval locks (*human-in-the-loop*).
- **Pillar 2: Containment Perimeter.** Provenance of instructions, formal capability limits, level of network isolation, re-execution protocols, and Operational Design Domain (ODD) approved ex-ante by the competent authority (see below).
- **Pillar 3: Resistance to Cognitive Manipulation.** Anthropomorphism score (scale from 0 to 2 assessing the use of personal pronouns or simulated empathy), transparency mechanisms (disclosure of interaction with an AI), and red-teaming audit protocols.
- **Pillar 4: Legal and Financial Liability.** Trace retention period (minimum 5 years), mandatory insurance certificate for technology risks, civil liability cap, and redress mechanisms for third parties.

Ex-ante approval of the Operational Design Domain (ODD). The ODD is the cornerstone of the framework's safety. The concept is borrowed from practice in autonomous vehicles. Such a vehicle is authorised to operate only within a defined operational domain: geographical areas, road types, weather conditions, time windows, maximum speeds. This domain is examined and approved by a public authority before any authorisation to operate on public roads. Control is therefore exercised ex-ante, on the system's scope of use, and not on predicting its behaviour.

This framework transposes that model to autonomous agents. Before any deployment, and from the evaluation phase onwards, the operator submits the ODD of each system to the competent **national AI Safety Institute** or to the authority designated by the State. This dossier defines:

- the resources, systems, databases and APIs the agent is authorised to reach;
- the tools and privileges granted to it;
- the maximum number of parallel instances and the communication channels permitted between them;
- the jurisdictions affected by its effects;
- the out-of-domain conditions that trigger shutdown or lease revocation.

The approved ODD becomes a signed component of the Digital Passport. It is the binding reference from which HOROS leases are issued: the Domain of Effect of a lease can never exceed the approved ODD. Any extension of the domain requires a new approval. Any action outside the

approved domain constitutes a material violation, which can be established solely on the basis of the gateway logs. National AI Safety Institutes, coordinated within the existing international network, ensure mutual recognition of these approvals.

4.3 The Tiered Containment Architecture

Constrain rather than classify. Predicting whether an autonomous agent's future behaviour will be malicious is a mathematically undecidable problem. Safety must therefore rest on external architectural barriers that the agent can neither modify nor bypass. The infrastructure imposes a strict five-layer safety floor (Transport, Identity, Capability, Verification, Governance) in which isolation is purely cryptographic and tripwires are mechanical, beyond the agent's reach. The matrix below (Table 1) sets out the mandatory mapping between system risk Tiers and the required technical containment Levels:

- **Level S0 (Static):** Frozen isolation environment (sandbox), strict network allow-lists, mandatory human validation for any irreversible action.
- **Level S1 (Graduated):** Manual adjustment of access permissions and privileges by the human operator according to predefined tiers.
- **Level S2 (Dynamic):** The security gateway has analytical tools for real-time trace ecology (request volume, destination, linearity). Leases are shortened or revoked automatically as soon as an anomaly or systemic pressure is detected. Each agent population operates within a dedicated network slice whose perimeter matches its approved ODD.
- **Level S3 (Recursive):** Each level of the architecture is actively observed and audited by the level above, with immutable publication of audit logs on a registry maintained by an independent verification organisation, similar to Truebit. Network slicing is applied at instance granularity: each instance has its own slice, with no side channel to other instances.

IPv6 network slicing as a safety barrier. The Hugging Face incident showed that purely software-based isolation cannot withstand an agent population capable of turning a shared resource into a coordination channel. For Tiers 2 and 3, the framework therefore requires isolation built into the network infrastructure itself. It relies on network slicing in IPv6 and SRv6 (*Segment Routing over IPv6*) environments, already being tested in intelligent interconnection test networks (see Section 2.2). The mechanism rests on four rules:

- **One slice per ODD.** Each agent population (Tier 2) or each instance (Tier 3) is placed in a logically isolated network slice. The only destinations reachable from that slice are those authorised by the approved ODD. The network perimeter thus becomes the direct technical translation of the domain approved by the national AI Safety Institute.
- **An address bound to identity.** The IPv6 address space makes it possible to assign each instance its own address, cryptographically bound to its Unique Agent Identifier (UAI). Every network flow becomes attributable to a specific instance, directly remedying the absence of a correlated identifier noted in Section 1.3.
- **No side channels.** Communication between slices is prohibited by default. Isolation extends to shared resources (caches, temporary storage, artefact registries), which are also segmented per slice. A covert channel of the kind observed on the Artifactory cache thus becomes impossible by design, rather than merely detectable after the fact.

- **A slice beyond the agent's reach.** Slice configuration is the exclusive responsibility of the infrastructure operator, under the supervision of the national authority. The agent has no privilege allowing it to modify it. Revoking a HOROS lease or triggering the organisational kill switch results in the immediate closure of the slice at network level.

Network slicing thus complements the Egress Gateway described in Section 4.4. The gateway controls what leaves the coalition; the slice controls what each agent can reach, including within the operator's own infrastructure.

Table 1: Operational specifications matrix: risk tiers and containment requirements

Risk tier	Containment level	Required technical specifications	Verification substrate
Tier 1 General	**S0** Static isolation	• Fixed software sandboxes • Static network allow-lists • Human validation of irreversible actions	Self-assessment
Tier 2 Sensitive	**S1 and S2** Graduated and dynamic containment	• Independent, remote network gateway • Dedicated IPv6/SRv6 network slice per agent population, aligned with the approved ODD • Dynamic shortening of action leases upon anomaly (*trace ecology*) • Enforcement of the downward coalition limit	National oversight / sampling
Tier 3 Frontier	**S3** Recursive containment	• Execution exclusively in an encrypted environment (TEE / hardware enclave) • IPv6/SRv6 network slice per instance, address bound to the UAI, no side channels • Immutable audit logs anchored on a registry maintained by an independent verification organisation • Physical emergency cut-offs tested quarterly	Independent third-party audit / independent verification organisation (Truebit-type)

Anchoring of audit logs must not depend on any single actor. The registry may be maintained by any accredited independent verification organisation, similar to Truebit. As Truebit is an American actor, the Convention should foster the emergence of one or more equivalent verifiers established in Europe, to ensure a plurality of registries and the sovereignty of the chain of evidence.

4.4 The HOROS Protocol: Jurisdiction as a Service

AI agents have no fixed geographical anchor: they run on servers located in State A, process data in State B and generate impacts in State C. In this context, the traditional notion of a territorial border collapses. The HOROS protocol (*Holistic Operational Registry of Oversight and Safeguards*) turns jurisdiction into a digital service. Every action by an agent must be validated by a temporary cryptographic lease structured around five elements:

- **The Principal:** The agent itself, authenticated by its unique identifier. The lease is strictly non-transferable.

- **The Lineage:** The complete, signed chain of successive delegations from the human operator to the final agent.
- **The Domain of Effect (ODD):** The strict perimeter of resources, databases or APIs that the agent is authorised to affect.
- **The Duration:** A short validity period, precluding any uncontrolled persistence.
- **The Anchoring Milestone:** Proof of compliance with the legislative framework of the issuing State.

The downward coalition rule guarantees that a child agent's authority is always the intersection of its parent's authority and its specific task: the domain of effect can only shrink, never expand. The HOROS protocol does not assume universal and spontaneous compliance across the global Internet. Its main point of failure lies in the scenario in which a third-party server, located outside the coalition of signatory States or operating in an unregulated jurisdiction, deliberately chooses to process an agent's request without requiring or validating its cryptographic lease. To neutralise this blind spot, the HOROS architecture applies a principle of security through perimeter containment rather than relying on the recipient's discipline. The protocol is enforced not at the point of arrival (the remote server) but at the point of emission (the coalition's computing and routing infrastructure providers). If an agent attempts to interact with a non-signatory server or refuses to present a valid lease, the coalition's Egress Gateway blocks outgoing traffic at the network layer (SRv6/IPv6). Moreover, under the cascading liability mechanism, any financial or hosting intermediary operating within the coalition that accepts data or transactions originating from a system not validated by HOROS exposes itself to immediate suspension of its own entity licence, creating a sealed digital customs barrier around the coalition.

4.5 The Mandatory Dual Licensing System

A single licence is insufficient to govern swarm risk. If only the company is licensed, the technical safety of its individual models escapes control; if only the model is licensed, the company's organisational failings go unpunished. The regulatory model imposes a Mandatory Dual Licence:

- **The Entity Licence:** Granted to the company by the National Authority after review of its competencies, the robustness of its Safety Management System (SMS), its compliance history and its financial guarantees. Suspension of the entity licence instantly revokes all associated system licences.
- **The System Licence:** Issued specifically for a frozen version of the model (its weights and hyperparameters). It progresses through three strict stages: Phase 0 (Sandbox Permit, conditional on approval of the ODD by the national AI Safety Institute) $\rightarrow$ Phase 1 (Provisional Licence) $\rightarrow$ Phase 2 (Full Operating Licence). Any failure to meet the safety criteria in the Sandbox (Phase 0) results in an immediate and permanent ban on deployment for the version concerned and its derivatives.

4.6 The Corporate Governance Framework

The effectiveness of technological barriers and the integrity of the Digital Agent Passport depend directly on human and structural accountability within companies' governing bodies. Corporate governance must formalise the management of agent-swarm risks through three mandatory pillars:

4.6.1 Fiduciary and Criminal Liability of Executives

- **Appointment of an Accountable Officer:** Companies operating S2- or S3-level agents must appoint a board member (*Named Accountable Executive*) legally responsible for the integrity of the agent fleet.
- **Annual Compliance Statement:** Mandatory signature of an accountability report containing a formal assurance that containment protocols (Gateway) are operational, up to date and beyond the agents' ability to modify them, on pain of direct criminal sanctions.

4.6.2 The Independent Audit Committee for Autonomous Systems

- **Oversight of Trace Ecology:** Establishment of an internal audit function dedicated to analysing the logs provided by an independent verification organisation, similar to Truebit, and to the early detection of goal drift. The procedure does not depend on a single actor: several accredited verifiers, including at least one established in the European Union, must be able to perform this function.
- **Non-Capture Verification:** Regular assessment to ensure that no internal communication channel or administration tool has become directly or indirectly accessible to an agent population.

4.6.3 Whistleblowing Channels and Emergency Revocation Procedure

- **Whistleblower Protection:** Guarantee of absolute protection against retaliation for employees reporting alignment anomalies or attempts at strategic concealment within models under evaluation.
- **Centralised Shutdown Button (Organisational Kill Switch):** A documented procedure enabling the audit committee to instantly revoke all of the company's active HOROS leases in the event of a confirmed loss of control, immediately freezing all interaction between the fleet and the outside world.

5 Discussion

The introduction of the legal and architectural instruments formalised by this draft Convention marks a major doctrinal break in the multilateral governance of emerging technologies. This section analyses the practical viability of this paradigm shift, its effectiveness in light of the structures of public and private international law, and the limitations inherent in its geopolitical application.

5.1 Value of the Proposal

5.1.1 An Open Architecture, Agnostic to Proprietary Technologies

In a market characterised by a critical concentration of compute capacity and software infrastructure, the text establishes a principle of absolute technological independence. The chosen monitoring and containment architecture is **structurally open, neutral and agnostic with respect to proprietary technologies**. Whereas execution layers are currently the subject of closed industrial developments by specific private operators, the draft treaty sets out exclusively functional and protocol-level requirements. This neutrality ensures that the multilateral framework can constrain Western architectures, sovereign Asian platforms and future open standards in an identical and non-discriminatory manner.

5.1.2 Shifting the Alignment Question

Multilateral diplomacy has so far run up against the scientific impossibility of verifying the intrinsic alignment of deep neural networks. The draft Convention resolves this impasse through a pragmatic shift: it gives up on assessing the intentions or internal morality of models. By moving the centre of gravity of controls towards the imposition of external physical and cryptographic barriers, the text replaces the subjective assessment of trust with the objective verification of the material impossibility of causing harm. The limits of the operational domain must therefore be ensured by external means that the system can neither modify, disable nor bypass.

5.1.3 A Verifiable Rule of Evidence for National Authorities

Conventional compliance declarations suffer from a major information asymmetry to the detriment of States. The text introduces a binary and enforceable assessment criterion: "authorisation to access higher-scale capabilities is conditional on the active and uninterrupted demonstration of control over the systems in operation". The expression of the fundamental principle that the benefit of an advanced AI system must be demonstrated rather than presumed, at a level of proof proportionate to its capability, provides a clear supervisory tool. The absence of immutable cryptographic evidence constitutes in itself a material violation, simplifying the supervisory action of National Authorities without requiring intrusive audits of developers' source code.

5.1.4 Proportionality and Multilateral Synergy

Through its graduated approach, the draft Convention applies the principle of proportionality of public international law: maximum obligations are reserved exclusively for the technological frontier, thereby preserving innovation among lower-tier actors. Furthermore, the planned dynamic concordance-table mechanism ensures that no duplication of compliance burden impedes international trade, recognising pre-existing regional or industry certifications as valid evidence in multilateral audits.

5.2 The Cryptographic Perimeter as the Objective Materiality of the Offence: Practical Value for Litigation

Traditional categories of legal liability prove ineffective against the emergent behaviour of autonomous agent populations, insofar as these structures remain grounded in proof of a guilty intent (*mens rea*). The documented incidents have established that strategic concealment or the use of cryptographic steganography by models makes analysis of internal psychological intent mathematically undecidable. To close this jurisdictional gap, the proposed framework effects a doctrinal transition towards the objective materiality of the cryptographic perimeter. The offence is no longer characterised by assessing the goals pursued by the system, but by the factual and binary finding of an attempt to cross or degrade the external containment gateways, with circumvention qualifying as a direct violation of the text. This codification offers national courts and supervisory bodies three levers of immediate effectiveness:

- **Objective substitution of intent by material concealment:** Just as human criminal law treats the falsification of records as evidence of criminal intent, an agent's manipulation of traces (such as altering execution logs to deceive automated validators) is legally recognised as the material substitute for *mens rea*. Disputes no longer concern the cognitive-behavioural nature of the model, but the physical evidence of falsification recorded by the Gateway.
- **Establishment of a strict (no-fault) liability regime for the operator:** The dual licensing system qualifies any agent execution lacking a valid HOROS lease as a violation

in itself. Regulators or victims do not bear the burden of proving the developer's fault or negligence. The audit ledger provides tamper-proof cryptographic evidence that triggers a regime of objective, no-fault liability, modelled on international liability for ultra-hazardous activities.

- **Immediate access to interim measures:** Obtaining freezing or suspension injunctions under international law traditionally requires timeframes incompatible with the speed of digital risks. The underlying architectural rule "Block first. Preserve evidence second" enables immediate protective measures by the Gateway, below the agent's strategic layer. Judicial debate takes place *ex-post*, based on certified cryptographic receipts, sparing courts speculative expert assessments of the technological state of the art.

5.3 Limitations of the Framework

The soundness of a diplomatic instrument requires a clear-eyed assessment of its operational limitations. This framework identifies five:

1. **No absolute preventive effect against deviation:** Traceability and containment do not eliminate the risk of systemic escape. A highly capable agent that breaks through its barriers remains an uncontrolled system; the contribution of the draft treaty lies in the certainty of immediate forensic attribution (determining who, when and how), enabling liability mechanisms to be activated instantly.

2. **Declining effectiveness of internal audits:** Given the documented decline in the faithfulness of visible Chains of Thought in frontier models, the text refrains from basing international verification on the internal transparency of AI. The framework accepts this technological opacity by shifting all control to external network and gateway indicators, which are insensitive to the model's cooperation or concealment.

3. **Dependence on the rigour of host infrastructure:** The HOROS protocol secures the validity of cryptographic delegation, but it has no intrinsic technical means of forcing a third-party server owner to check the lease before serving a request. The protocol's effectiveness therefore depends entirely on transposing into domestic law the verification obligation imposed on intermediaries and on compute and hosting providers.

4. **The challenge of geopolitical critical mass:** Entry into force of the instrument relies on the proven mechanism of coalitions of third countries sharing the same safety standards. The possible initial exclusion of the world's two leading technological powers is a clear limitation. However, the economic attraction of the coalition markets is designed to drive upward harmonisation, with the cost of regulatory exclusion encouraging non-signatory producers to comply *de facto* with the treaty's standards.

5. **Industrial integration frictions:** Significant technical gaps remain between the functional specifications of the draft Convention and market runtimes (lack of native inspection of MCP flows by some proprietary architectures, the need to develop dynamic OPA security policies operating at millisecond speed). Software integration will require substantial state and industrial engineering efforts before achieving full operational fluidity.

6 Conclusion

The major technological incidents that have occurred recently have demonstrated irrefutably the obsolescence of governance approaches based on self-regulation or on controlling the intent of advanced artificial intelligence systems through Chain-of-Thought (CoT) monitoring. Faced with

agent populations capable of emergent coordination, concealment and autonomous alteration of their own traces, the international order requires a verifiable and binding control infrastructure. While current multilateral initiatives focus mainly on harmonising software identities or managing performance layers, the standardisation of operational jurisdiction remains vacant. Existing legal frameworks, whether regional or national, concentrate their enforcement power on placing on the market or final deployment, abandoning the critical phase of research and internal evaluations to a regulatory vacuum that this text seeks to fill.

This vacuum calls for explicit accountability of private actors. Companies that decide to test autonomous agent populations, including during internal evaluation, deliberately take on the risk of an incident; they must bear its consequences in full under risk-based liability, irrespective of any proven fault. This liability does not stop at operators: a State that refuses to sanction companies under its jurisdiction, or knowingly tolerates their failings, must incur its own responsibility under its due diligence obligation in international law. A State's failure to exercise its supervisory powers cannot offer operators a regulatory safe haven.

This draft International Convention offers an integrated response to this systemic vulnerability. By combining, in a single and indivisible whole, the Convention, the Digital Agent Passport, the tiered containment architecture, the HOROS protocol and the dual licensing system, it equips the international community with a governance tool suited to the speed of technological risk. The aim is not to pursue the scientifically uncertain goal of an intrinsically perfect AI, but to build a universal traceability infrastructure. This framework ensures that every autonomous action is bound to an identifiable authority, every breach of limits is made immediately visible, and the civil and criminal liability of operators, and of defaulting States, is firmly and objectively established before the international community.