THOUGHT LEADERSHIP SERIES

The Post-Market Paradox

Part 4 of 4: Executing Continuous Compliance Monitoring Across Cloud and On-Premises Deployments

April 2026

The EU AI Act fundamentally breaks the traditional software procurement model. Under Article 72, an AI vendor's legal responsibilities do not conclude at the point of sale. The EU AI Act mandates that Providers operate a continuous, active post-market monitoring system that tracks their AI system's real-world performance throughout its entire operational lifetime. However, this mandate creates a profound architectural conflict when AI systems are deployed securely on-premises. How does a Provider continuously audit an AI system that is locked behind a client's corporate firewall? This white paper, the final installment in our four-part series on autonomous AI auditing, explores the Post-Market Paradox and details the cryptographic architectural solutions required to resolve it.

1. The Obligation That Never Ends

As Part 2 of this white paper series established, AI models are statistically dynamic systems: they drift. Article 72 of the EU AI Act exists precisely because regulators understand this mathematical reality. The EU AI Act explicitly requires Providers to establish a continuous post-market monitoring system that proactively collects real-world performance data to ensure their AI systems remain safe, fair, and accurate long after initial certification [1].

For the Provider, this is not merely a customer service function; it is a strict legal obligation. If a high-risk AI system begins to exhibit dangerous statistical drift or catastrophic failure in the wild, the Provider is legally accountable for catching it and reporting it. To execute this, Article 72(2) explicitly states that monitoring systems must collect 'data which may be provided by deployers,' formalizing a mandatory, continuous feedback loop between buyer and seller [1]. The challenge is not understanding this legal obligation; the challenge is physically executing it across vastly different enterprise IT architectures.

2. The Architecture Divide: SaaS vs. On-Premises

The physical architecture of how an AI system is deployed dictates the entire compliance strategy. The market currently relies on three primary deployment models, each presenting radically different audit challenges:

- SaaS / Public Cloud API: The Provider hosts the AI system centrally on its own servers, and the Deployer accesses it via API. From an Article 72 perspective, this is structurally simple for the Provider: it has full visibility into the AI system's telemetry and can easily monitor output distributions. The critical compliance risk here is instead GDPR and data privacy, as the Provider is actively processing the Deployer's potentially sensitive prompt data on external servers, exposing the Deployer to data exfiltration and localization risks. Recent industry analysis of the AI-as-a-Service (AIaaS) market confirms that while hyperscalers attempt to mitigate this by offering localized cloud regions, the fundamental risk of exposing proprietary data to third-party API endpoints remains a primary barrier for highly regulated enterprises [2].
- On-Premises / Secure Edge: To protect trade secrets, intellectual property, or classified data, enterprise Deployers such as defense contractors, hospitals, and financial institutions frequently demand that the AI system be containerized and deployed deep inside their own highly secure corporate infrastructure. Recent empirical research on LLM deployment confirms that migrating from public cloud APIs to on-premises environments is a primary mechanism that organizations use to definitively control data location and mitigate privacy risks [3]. However, while this solves the Deployer's security concerns, the AI now operates entirely behind the firewall, completely air-gapped from the Provider.
- Internal Deployment (In-House Build): Enterprises that develop and deploy their own AI models strictly for internal use behind their own firewalls often mistakenly assume they are exempt from the regulation because they never placed the system on the market. However, recent legal analysis confirms that the act of putting an AI system into service internally fully triggers the scope of the EU AI Act. The enterprise legally becomes both the Provider and the Deployer simultaneously, forcing it to build its own Article 72 continuous monitoring infrastructure from scratch [4].

3. The Firewall Paradox: Auditing the Unreachable

The on-premises deployment model creates what we term the Post-Market Paradox, a direct conflict between two immovable compliance obligations:

- The Provider's Mandate: Under Article 72, the Provider is legally obligated to continuously monitor the model's performance in real time.
- The Deployer's Mandate: Under internal data governance policies, GDPR, and basic operational security, the Deployer's CISO is legally and contractually obligated to prevent any external exfiltration of internal data from behind the corporate firewall [5].

These two mandates appear structurally irreconcilable. A highly regulated enterprise will never permit a vendor's engineers to remotely extract live operational data across the firewall boundary. Conversely, the Provider cannot legally fulfill its Article 72 obligations if it is entirely blind to how its model is behaving in production. Conventional solutions, such as periodic manual audits, quarterly questionnaire-based assessments, or contractual SLAs, cannot resolve this paradox. As recent research from the Karolinska Institutet on post-market governance of adaptive medical AI highlights, manual and case-based reporting infrastructures are insufficient for properly assessing the safety and effectiveness of AI/ML devices, because algorithmic degradation is invisible to periodic, point-in-time checks [6].

4. The Solution: Autonomous Auditing at the Edge

Resolving the Post-Market Paradox requires an architectural innovation rather than a legal negotiation. Rather than attempting to pull raw operational data out of the Deployer's secure environment, the Provider must deploy the audit mechanism inside the firewall [7].

This is achieved through the deployment of lightweight autonomous auditing agents that run locally alongside the containerized AI model. This three-layer mechanism operates as follows:

1. Local Evaluation: The autonomous agent continuously monitors model performance, output distributions, and statistical drift indicators entirely within the Deployer's secure infrastructure. No raw data, user prompts, or PII ever leaves the perimeter.
2. Cryptographic Compression: The agent distills its behavioral findings into mathematically encrypted, anonymized compliance metadata. It generates statistical proofs of performance, completely stripped of the underlying sensitive context.
3. Secure Transmission: Only this compressed, encrypted compliance signal is transmitted back to the Provider. This provides cryptographic evidence that the model is functioning within defined safety thresholds, satisfying the Provider's Article 72 monitoring mandate without breaching the Deployer's data governance policies or triggering GDPR violations [7].

5. Board Action Items

As organizations operationalize their AI governance frameworks, executive leadership must ensure compliance strategies account for the physical architecture of their AI deployments:

- Map the Deployment Architecture: Identify every on-premises or edge AI deployment across the enterprise where Article 72 creates a monitoring blind spot between the Provider and the Deployer.

- Audit Vendor Contracts for the Paradox: Ensure SLAs explicitly address how the Provider will fulfill its continuous post-market monitoring obligations without violating internal data-exfiltration policies.
- Mandate Autonomous Auditing Agents: Specify in all future AI procurement contracts that vendors must provide locally deployable, cryptographically verified compliance monitoring agents as a standard architectural requirement for on-premises deployments.

Series Conclusion

The EU AI Act represents a seismic shift in corporate liability. As this four-part series has demonstrated, the sheer statistical complexity of AI systems renders traditional, human-led, point-in-time audits obsolete. From the Provider grappling with the impossible mathematics of black-box proof in White Paper Part 2, to the Deployer attempting to govern human-in-the-loop automation bias in White Paper Part 3, the regulatory burden is immense. To survive this new era of liability, organizations must structurally shift from manual compliance checks to the continuous, mathematically verified Continuous Auditing of AI Systems (CAAI).

As the latest 2026 regulatory mapping for AI providers explicitly concludes: if a provider cannot continuously demonstrate that the system's behavior remains within the boundaries assessed during the initial conformity assessment, and cannot instantly detect when it deviates, then the essential legal requirements for accuracy, robustness, logging, and post-market monitoring are unfulfilled by definition [8]. By embedding autonomous auditing directly into the infrastructure, whether in the cloud or at the secure edge, organizations can deploy transformative AI technologies while maintaining cryptographically verifiable regulatory compliance evidence.

Sources

1. EU AI Act (Regulation 2024/1689) - Article 72: Post-market monitoring by providers - https://artificialintelligenceact.eu/article/72/
2. V. Nikitiuk, AI as a Service in the EU Market: Navigating Business Challenges and Regulatory Compliance Through Industry Insights, Master's thesis, 2025, Uniwersytet im. Adama Mickiewicza w Poznaniu and European University Viadrina Frankfurt - https://www.researchgate.net/publication/401622001_AI_as_a_Service_in_the_EU_Market_Navigating_Business_Challenges_and_Regulatory_Compliance_Through_Industry_Insights
3. T. Paloniemi, M. Setälä and T. Mikkonen, 'Porting an LLM based Application from ChatGPT to an On-Premise Environment,' 2025 IEEE/ACM 22nd International Conference on Software and Systems Reuse (ICSR), pp. 78-83 - https://www.computer.org/csdl/proceedings-article/icsr/2025/261700a078/27t2lTZRwoE
4. M. Pistillo (2025), arXiv:2512.05742 - https://arxiv.org/abs/2512.05742
5. J. Mokander and L. Floridi, Minds and Machines 31, 323-327 (2021) - https://link.springer.com/article/10.1007/s11023-021-09557-8
6. F. Afdideh, M. Astaraki, F. Seoane and F. Abtahi (2026) - https://api.semanticscholar.org/CorpusID:286770412
7. M. Minkkinen, J. Laine and M. Mäntymäki, DISO 1(3), 21 (2022) - https://link.springer.com/article/10.1007/s44206-022-00022-2
8. L. Nannini, A. L. Smith, M. J. Maggini, E. Panai, S. Feliciano, A. Tiulkanov, E. Maran, J. Gealy and P. Bisconti (2026) - https://arxiv.org/abs/2604.04604