The Inference Gap: Test-Time Compute Scaling and the Limits of Training-Centric AI Governance

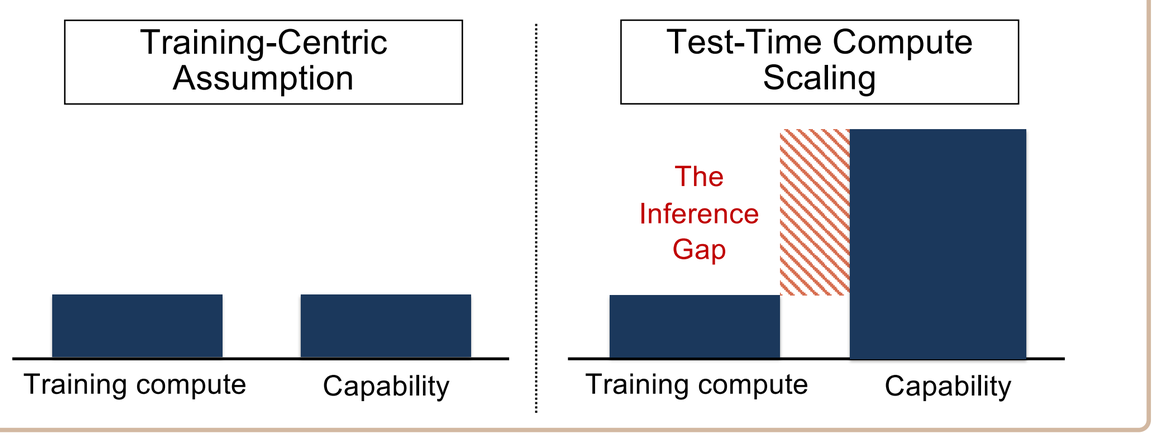
Luiz Felipe Caparelli Piochi[1,*]
[1] Université de Lorraine, Inria, F-54000 Nancy, France
* Correspondence to: *luiz-felipe.caparelli-piochi@inria.fr*

1. The Inference Gap

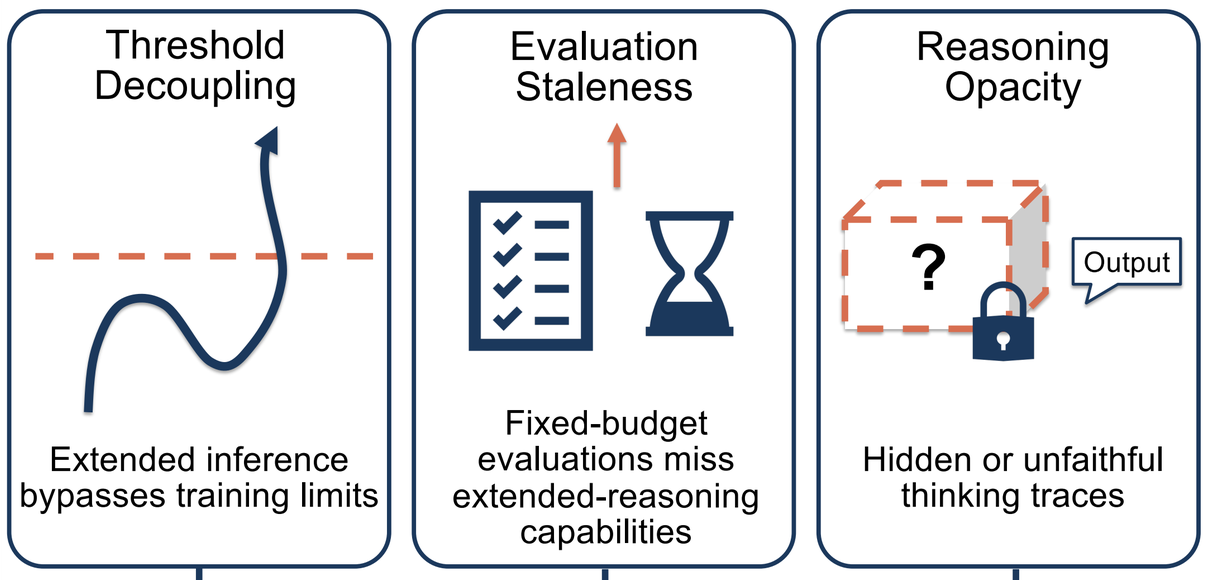
Contemporary AI governance assumes capability is determined at training time. The EU AI Act applies its systemic risk tier to models trained above 10^{25} FLOPs. The revoked US Executive Order 14110 set a reporting trigger at 10^{26} FLOPs. Both treat training compute as a sufficient proxy for deployment capability.

Test-time compute scaling breaks this assumption. The same model weights produce qualitatively different outputs depending on inference budget. Chain-of-thought prompting, best-of-n sampling, and tree-search methods can dramatically improve performance without changing weights, exposing **the inference gap**.

DeepSeek-R1, for example, was trained at approximately 3.5×10^{24} FLOPs, yet it still matched OpenAI o1's reported performance. Training compute and deployment capability are no longer reliably co-predictive, but current governance standards still lag behind.



2. Three Governance Gaps, Three Proposed Reforms



Threshold Decoupling. Sub-threshold trained models paired with extended inference can match frontier capability outside enhanced regulatory scrutiny. The capability profile under extended inference is a joint function of weights and inference configuration. DeepSeek-R1 instantiates this pattern.

Evaluation Staleness. Pre-deployment evaluations conducted at fixed inference budgets cannot characterise the capability profile of a model later deployed with higher inference budgets. Smaller models with extended reasoning can outperform models fourteen times larger at standard settings. Safety thresholds calibrated to non-reasoning baselines misevaluate extended-reasoning deployments.

Reasoning Opacity. Hidden scratchpads in closed-source models are architecturally inaccessible to external auditors. Visible chain-of-thought, where exposed, has been shown empirically to diverge from the causal factors that determine model outputs. Auditability requires both access and faithfulness, and neither holds without explicit governance requirements.

Inference Budget Declaration

What: Declare the maximum per-call reasoning budget configurable at deployment.

Verification: Signed API specifications and audit-grade logs.

Multi-Budget Capability Evaluations

What: Run dangerous capability evaluations at the maximum deployment inference budget.

Verification: Multi-budget benchmark calibrated to the deployed ceiling.

Reasoning Trace Logging

What: Log full reasoning traces for high-risk applications, auditor access on request

Verification: Disclose latent reasoning and process-reward training that erode faithfulness.

Scope and Limits

API-mediated deployments. Inference budget declaration applies where providers control the infrastructure. Open-weight models require model-card disclosure instead.

Distillation not covered. Sub-threshold students can inherit teacher capability without inference amplification. A separate decoupling vector addressed by provenance disclosure.

Faithfulness is open. Trace logging delivers value only if logged content reflects underlying computation. Mechanistically faithful chain-of-thought remains an active research area.

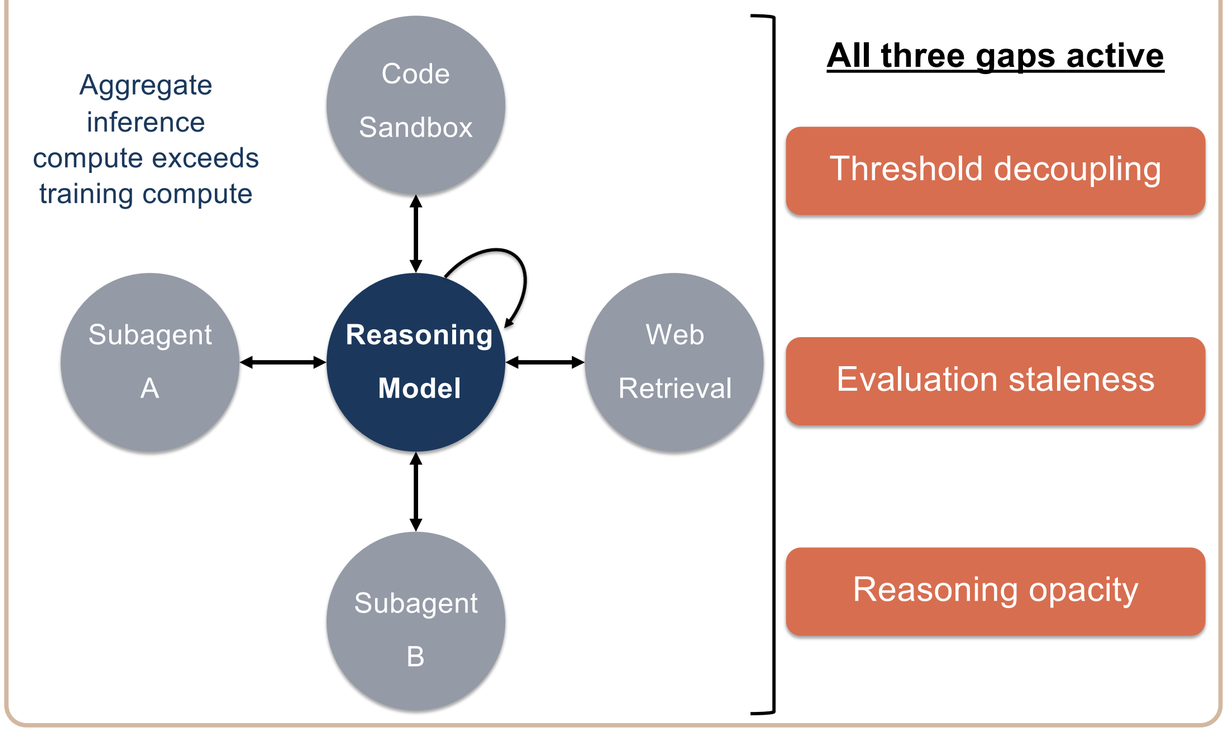
Jurisdictional scope. Specified against the EU AI Act. UK and US AI Safety Institute protocols share the same architecture and admit the same reforms.

3. Agentic Compounding

All three gaps amplify when reasoning models are deployed as agents. A single agentic task can involve hundreds of model calls, each with extended reasoning. Aggregate inference compute can far exceed the underlying model's training budget.

Single-call evaluations miss behaviours that emerge from composed reasoning, tool use, and retrieval. Distributed reasoning traces spread across calls, instances, and tool outputs are not systematically retained.

Consider a cybersecurity agent built on a sub-threshold base model with code execution and web retrieval. In isolation at standard budget, it scores below dangerous-capability thresholds. Deployed as described, it iteratively enumerates attack surfaces, tests payloads, and retrieves current CVE disclosures. All three gaps activate simultaneously.

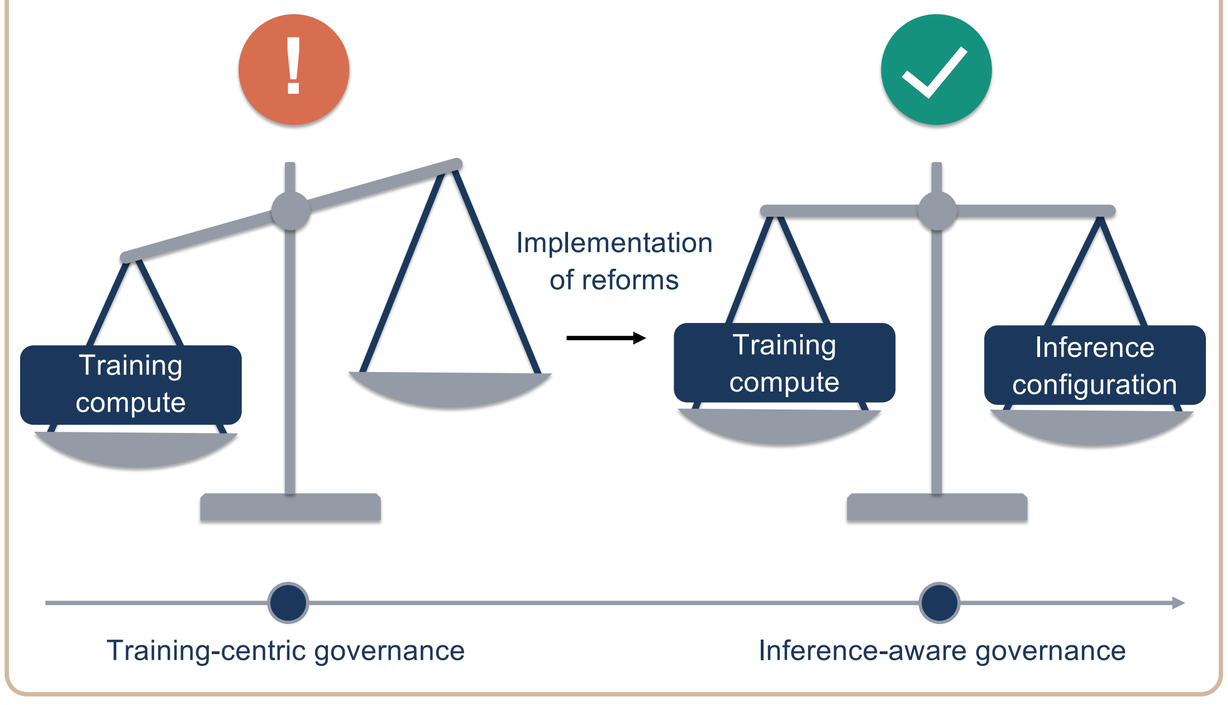


4. Takeaways

The training-centric governance paradigm has been superseded by deployment architecture. Inference configuration is now a determinant of capability alongside model weights, and it must become a governable parameter in its own right. Treating the trained artifact as the sole object of oversight leaves the deployed system out of scope.

The proposed reforms do not require resolving open problems in alignment research or interpretability. Inference budget declarations, multi-budget capability evaluations, and reasoning trace logging can each be specified as technical annexes to the EU AI Act GPAI Code of Practice and adopted by national AI Safety Institutes through evaluation protocol updates. None of the three depends on a research breakthrough.

The models and deployment patterns that expose these gaps are in widespread production use, not future systems. Governance frameworks designed for an earlier paradigm need targeted technical updates now.



Acknowledgments. *The author thanks colleagues at* ***Inria*** and ***Sciences Po****, as well as* ***Ana Clara Tardem****, for the discussions that shaped this work.*
Key references. *Ord 2025, Inference scaling reshapes AI governance (arXiv). Sastry et al. 2024, Computing power and the governance of AI (arXiv). Snell et al. 2025, Scaling LLM test-time compute optimally (ICLR). Guo et al. 2025, DeepSeek-R1 (Nature). European Parliament and Council 2024, EU AI Act, Regulation 2024/1689.*