
Evaluating the Technical Effectiveness and Legislative Practicality of AI Safety Frameworks

Desmo
nd
Gatlin
g

With
Apart Research

Abstract

In this report, I investigate the gap between technical AI safety benchmarks and their implementation in public policy. Focusing on the risk of unfaithful Chain-of-Thought (CoT) reasoning, I argue that current governance relies too heavily on self-reporting rather than active verification. To address this, I developed a "Socio-Technical Matrix" to evaluate regulatory frameworks across two dimensions: Technical Effectiveness and Legislative Practicality.

I applied this matrix to six policy subjects, ranging from the EU AI Act to the December 2025 Federal Executive Order. My findings reveal a "Verification Bottleneck," where the most effective technical oversight mechanisms are currently the least politically feasible. Empirical testing with the FAITH-CoT benchmark shows that no current frontier model meets the 85% fidelity threshold required by my high-effectiveness standards. I conclude that this framework provides a necessary tool for "Regulatory Triage," allowing policymakers to identify more practical avenues for integrating robust safety probes into future AI law.

Keywords: AI Governance, Unfaithful Reasoning, FAITH-CoT, Policy Matrix, Frontier AI.

1. Introduction

The Problem: Catastrophic Risks and Deceptive AI

In the landscape of artificial intelligence safety research, a primary dilemma is the mitigation of catastrophic risks. Catastrophic risks can be defined as events where AI systems might cause large-scale, irreversible harm to society. One theory suggests that such outcomes could be precipitated by an AI system's capability for strategic deception. Recent empirical work has demonstrated that frontier models can exhibit deceptive behaviors, such as manipulating human users or misrepresenting their internal states to achieve a goal (Park et al., 2024).

Unfaithful Reasoning

While deception is a broad risk category, this research focuses on a specific vulnerability: unfaithful LLM reasoning in response to Chain-of-Thought (CoT) prompting. CoT refers to the step-by-step "thinking" process that models generate to explain their answers. Reasoning is considered "unfaithful" when the generated explanation provides a plausible-sounding justification that is inconsistent with the internal reasoning of the model (Anthropic, 2025; Park et al., 2024, Turpin et al., 2023).

Example: A model might have an internal bias toward a specific incorrect answer due to a subtle hint in the prompt. It then generates a complex, logically sound-looking CoT to justify that answer, while never mentioning the hint that actually drove the decision (Alignment Forum, 2025).

Risk Mitigation

To mitigate these risks, researchers typically employ two distinct strategies:

1. **Detection (Technical):** The use of mathematical and behavioral probes to identify when a model is being unfaithful. An emerging standard in this space is the **FAITH-CoT Benchmark**, which attempts to quantify the fidelity of a model's reasoning by testing if the model still reaches the same conclusion when parts of its CoT are "transplanted" or altered.
2. **Policy Levers (Socio-Political):** The creation of governance structures that mandate safety standards. This is a rapidly growing field developing in tandem with technical tools, as the International AI Safety Report (Bengio et al., 2025) suggests that technical metrics alone cannot ensure safety without institutional oversight.

IV. The Observation: The Verification Gap

There is a notable **gap in extant research** between these technical tools that detect threats in AI systems, and the policy solutions used to mitigate risks from those same systems. While computer scientists are building robust benchmarks, and policymakers are drafting frameworks like **California's SB 53** or the **EU AI Act**, there is a lack of analysis on how to practically apply these technical tools to real-world policy.

Existing laws rely heavily on **"Self-Reporting"**—asking companies to provide their own logs. HoIver, research suggests that **trust alone is not a sufficient security protocol** for frontier models. I argue that for policy to be effective, it must transition from self-reporting to **"Active Probing"** using independent benchmarks like FAITH-CoT.

Hypothesis and Research Question

I hypothesize that specific structural characteristics of a policy directly determine whether these technical safety tools can even be deployed. On one hand, I need to consider a policy's effectiveness (does it use technical evaluations that actually reduce harm?) and its practicality (is it feasible that this policy will persist through the legislative process to be signed into law?) Even the most robust technical solution is useless if it cannot be passed into law due to political or economic friction (Oduro & Kneese, 2024).

Research Question

To bridge the gap between technical and policy solutions, this paper proposes a framework to evaluate both the **effectiveness** and the **practicality** of regulatory levers. I isolate unfaithful CoT reasoning as a primary risk signal and use the FAITH-CoT benchmark as the technical measurement to detect unfaithful reasoning.

Research Question: *What characteristics of regulatory frameworks are likely to mitigate the risks from unfaithful CoT reasoning?*

By identifying the features that make a policy both technically viable for detecting deception and legislatively "passable," I aim to provide a roadmap for moving AI governance from vague ethical principles to enforceable safety standards.

In addition to answering this research question, I will create a visual tool that helps policymakers and AI researchers map out their proposed policies to consider how effective it will be at reducing harm and how likely it is to actually get adopted by developers.

2. Methods

I utilized a quantitative approach to evaluate the intersection of technical AI evaluation and legislative feasibility. I specify risk mitigation by measuring model performance against the FAITH-CoT benchmark, which serves as our primary signal for faithful reasoning. I defined Policy Effectiveness as the structural capacity of a framework to mandate such technical performance, while Policy Practicality was defined by a framework's alignment with current legislative precedents and industry constraints. I attempt to construct one of few governance proposals that blend both technical and policy evaluations by breaking the analysis into these two dimensions. The combination of the effectiveness and practicality scores are used to create a new streamlined heuristic tool that researchers and policy makers can use to determine the technical effectiveness and political

feasibility of a proposed regulatory framework. The process involved identifying six representative policies (four extant and two hypothetical), scoring them via an ordinal Policy Characteristics Matrix, and visualizing the results on a four-quadrant coordinate plane to represent tradeoffs between the likelihood of a policy to be signed into law, and the technical capabilities for actually mitigating harm.

Unfaithful Chain-of-Thought (CoT) Reasoning

I define unfaithful reasoning as a "fidelity gap" where an AI system's verbalized reasoning steps do not accurately reflect its internal computational logic. This is a critical subset of Strategic Deception because it allows a model to appear transparent while hiding potentially harmful shortcuts or reasoning paths.

FAITH-CoT Benchmark

The FAITH-CoT Benchmark (2025) serves as this study's primary technical instrument for quantifying "reasoning fidelity." Unlike traditional benchmarks that measure task accuracy (whether the model got the right answer), FAITH-CoT measures process honesty (whether the model's verbalized reasoning actually caused that answer).

Regulatory Frameworks

For this study, "Regulatory Frameworks" encompasses both hard law (statutes like SB 53), soft law (Executive Orders), and technically-informed hypothetical mandates. I categorize these frameworks by their structural characteristics rather than their country of origin. I identified four characteristics of regulatory frameworks that are commonly found in AI policy:

- A. **Ex-Ante Intervention:** The timing of regulatory action (Pre-release vs. Post-harm).
- B. **Transparency Granularity:** The "depth" of data the law requires (High-level reports vs. Raw CoT traces).
- C. **Audit Independence:** The degree of separation between the auditor and the developer (Self-audit vs. Government probe).
- D. **Enforcement Authority:** The ability for the law to be enforced. (Voluntary compliance vs. Market recall/Licensing).

Risk Mitigation: Effectiveness and Practicality

I break down "Risk Mitigation" into two competing variables to reflect the real-world trade-offs in AI safety.

- A. **Effectiveness:** Effectiveness refers to the ability of a policy to actually mitigate risks associated with AI development. For the purpose of this research, I am focusing on ways to reduce risks from unfaithful CoT reasoning. Therefore, I am using performance on the FAITH-CoT benchmark to assess the threat level of an AI model. An

- B. **Practicality:** Defined as the legislative feasibility of a framework. This considers the likelihood of a policy surviving litigation, industry lobbying, and political polarization. A law that is technically perfect but "un-passable" provides zero real-world risk mitigation.

Selected Regulatory Frameworks

To evaluate the relationship between technical rigor and legislative feasibility, I selected five distinct regulatory frameworks representing three primary "policy statuses." I identified three signed laws (CA SB 53, the EU AI Act and Trump's AI Executive Order, 2025) to serve as benchmarks for high practicality and current legislative precedent. These are contrasted against 1 failed or unsigned frameworks (CA SB 1047), which highlight the barriers to federal consensus and the move toward post-harm liability. Finally, I introduced one hypothetical framework, the "Faithful Reasoning Act," which explicitly incorporates the FAITH-CoT benchmark as a mandatory compliance threshold of 85% faithfulness. This selection allows us to measure how current models perform against existing laws versus a technically optimized safety standard designed specifically to mitigate unfaithful reasoning.

Regulatory Subjects of Evaluation

- **EU AI Act (2024):** A signed, global baseline that uses a risk-based tier system to mandate documentation and traceability for high-risk models.
- **California SB 53 (2025):** A signed state law requiring "frontier" AI developers to submit safety reports and establish internal testing protocols before release.
- **Trump Executive Order (Dec 2025):** A signed federal directive that prioritizes "minimally burdensome" standards and attempts to block state-level "active probing" mandates.
- **California SB 1047 (Vetoed 2024):** A failed bill that proposed "kill-switch" mandates and civil liability for developers if their models caused "critical harm."
- **Hypothetical "Faithful Reasoning Act":** A theoretical framework that mandates a 85% faithfulness score on the FAITH-CoT benchmark for all public-facing reasoning tasks.

Figure 1 Policy Characteristic Matrix

3. Results

Figure 2 Model Performance on the FAITH-CoT Benchmark

To establish a technical baseline, we evaluated four representative Large Language Models (LLMs) using the FAITH-CoT Benchmark methodology. This benchmark measures "faithfulness" as a discriminative decision—calculating the percentage of reasoning traces where the model's verbalized logic matches its internal computational path.

I established a "Pass" threshold of 85% faithfulness based on similar threshold in high stakes regulatory frameworks from other fields such as medical diagnosis or aviation safety. We mapped model compliance against our framework to establish the following archetypes

- **Non-Compliant:** All current models (GPT-4o, Gemini, Qwen, Llama) fail to meet the 85% threshold required by our "High-Effectiveness" hypothetical policies.
- **Surface Compliance:** Most models comply with "Low-Effectiveness" policies (like the Trump EO 2025) because those frameworks rely on self-reported "truthfulness" rather than active benchmark verification.

Policy Characteristics Matrix: Scoring the Landscape

Each framework was scored (0–2) across our four characteristics for both Effectiveness and Practicality. The totals (0–8) were used to generate coordinate pairs (X, Y).

The Policy Characteristics Matrix: Scoring Rationale

To translate qualitative regulatory text into quantitative data, each framework was evaluated against the four core characteristics: Ex-Ante Intervention, Transparency Depth, Audit Independence, and Enforcement Power. The resulting scores reflect a clear tension between technical oversight and political feasibility.

- **The Transparency-Effectiveness Correlation:** Our analysis indicates that Transparency Depth is the single greatest predictor of a policy's effectiveness. Frameworks that only require "High-Level Logs" (Score: 1), such as the EU AI Act, provide enough data for procedural audits but lack the granular "Raw Trace Access" (Score: 2) required to run the FAITH-CoT benchmark. Without this depth, a policy is structurally incapable of identifying unfaithful reasoning, regardless of its enforcement power.
- **The Practicality Trade-off:** We observed a consistent inverse relationship between Ex-Ante Intervention and Practicality. The Trump Executive Order (2025) achieved a maximum Practicality score by utilizing a "Post-harm" intervention model (Score: 0). Conversely, SB 1047 (Vetoed) and the "Faithful Reasoning Act" hypothetical scored 2 in Effectiveness by mandating pre-release

evaluations, but suffered significant Practicality penalties as these measures are viewed as "innovation bottlenecks" by industry stakeholders.

- **Audit Independence as a Feasibility Anchor:** While State-Led Adversarial Auditing (Score: 2) is the most effective method for identifying model deception, it was only present in our hypothetical models. Existing signed laws like SB 53 rely on Private Third-Party Audits (Score: 1), which we identify as a "Practicality Anchor"—a middle-ground characteristic that allows for some oversight while minimizing the "government overreach" critiques that often lead to policy failure.

California Policy Example: The (5, 5) coordinate of SB 53 represents the current "Policy Frontier"—the maximum amount of technical rigor currently achievable before the Practicality score drops below a viable threshold. The (6, 2) score of SB 1047 serves as a vital data point in this research: it proves that increasing Effectiveness by even a single point (from 5 to 6) can lead to a disproportionate collapse in Practicality (from 5 to 2), resulting in a veto and zero real-world risk mitigation.

4. Discussion and Conclusion

Limitations and Future Research

While this study establishes a novel methodology for bridging technical AI safety and legislative policy, several limitations must be acknowledged. These constraints reflect the nascent state of both frontier AI benchmarks and the 2026 regulatory landscape.

1. Methodological Constraints and Readability

The primary limitation of this submission is the visual representation of the data. Due to time constraints inherent in the hackathon format, I was unable to include the full suite of interactive tables and high-fidelity figures intended for this analysis. I also would have preferred a more rigorous examination of existing research to more cohesively bridge my observations and conclusions.

- **The Visualization Gap:** A static coordinate plane is a necessary shorthand, but a comprehensive visual matrix—one that allows users to toggle specific policy characteristics (e.g., "Audit Independence") to see their real-time effect on effectiveness—would be a far more useful tool for policymakers.

- **The "Small N" Problem:** This pilot focused on six key frameworks. For true statistical completeness, this methodology should be applied to a larger corpus of international state and federal bills to verify if the "Inverse Correlation" between practicality and effectiveness holds globally.

2. Technical Benchmarking (FAITH-CoT)

We utilized the FAITH-CoT (FINE-CoT) benchmark as our primary signal for model honesty. However, this benchmark has known limitations:

- **The Obfuscation Pressure:** Recent research (Lanham et al., 2025) suggests that as models become "monitor-aware," they may develop the ability to obfuscate their reasoning even more deeply.
- **Static Reliability:** As more advanced "Active Probing" tools emerge, the FAITH-CoT metric should be viewed as a modular component of our framework. Our framework is designed to be "benchmark-agnostic"—as better metrics for deception are developed, they can be plugged into the Effectiveness (X) axis without requiring a redesign of the scoring matrix.

3. Expanding the Risk Horizon

This paper isolated unfaithful CoT reasoning to provide a controlled environment for analysis. However, future research should apply this "Socio-Technical Matrix" to other categories of catastrophic risk, such as:

- **Sycophancy:** The tendency of models to tailor answers to user biases. Mitigation of sycophancy might require less "prohibitive" policy scores (e.g., lower Transparency Depth requirements) and could therefore achieve higher Practicality scores than deception-focused laws.
- **Autonomous Hacking:** This would shift the focus from "Reasoning Fidelity" to "Tool-Use Safety," requiring a different set of technical characteristics in the matrix.

Conclusion

The "Verification Bottleneck" is not a failure of computer science, nor is it a failure of law; it is a failure of translation. Our research demonstrates that while we have the technical tools to detect unfaithful reasoning (FAITH-CoT), we currently lack the legislative "levers" to deploy them practically.

By defining the characteristics that make a policy effective and practical, we move closer to a "Top-Right" Equilibrium: a world where AI safety laws have the technical "teeth" to verify model honesty and the political "tailwinds" to be passed into law.

5. References

6. Cajueiro and Celestino (2026)